Value of Hydrograph Characteristics or Single Discharge Observations in Hydrological Modeling

Dissertation
zur
Erlangung der naturwissenschaftlichen Doktorwürde
(Dr. sc. nat.)

vorgelegt der
Mathematisch-naturwissenschaftlichen Fakultät
der
Universität Zürich

von
Sandra Pool
von
Bregaglia GR

Promotionskommission
Prof. Dr. Jan Seibert (Vorsitz)
Dr. Daniel Viviroli
Prof. Dr. Christian Huggel

Zürich, 2018

Abstract

Discharge information is fundamental for many water management decisions. However, even in regions with a dense streamflow gauging network, most catchments are actually ungauged or poorly gauged, i.e., have no or only limited discharge data. In the absence of observed discharge, hydrological models are often used to simulate discharge time series. Hydrological models consist of parameters that quantify the storages and fluxes of water in a catchment. The value of these parameters can typically not be measured directly because the parameters usually represent several processes and integrate characteristics over the catchment scale. As a consequence, hydrological model parameter values have to be estimated by calibration, which consists of minimizing the difference between simulated and observed discharge using an objective error criteria (also called objective function). The need for accurate model simulations and the challenge of calibrating hydrological models and predicting discharge in data scarce catchments are the two foci of this PhD thesis.

The first part of this thesis was motivated by the need of accurate predictions of ecologically relevant streamflow characteristics (SFCs), which are ultimately needed for sustainable water management. In this thesis, the influence of the objective function, and as such the emphasis of certain hydrograph characteristics in calibration, on the estimation accuracy of multiple SFCs was explored. Calibration on commonly used objective functions (e.g. metrics based on the mean squared error) did in many cases not preserve a wide variety of SFCs and could result in unsatisfying estimates of SFCs. Directly including specific SFCs into model calibration could strongly improve their estimate, but generally resulted in an inadequate representation of other hydrograph characteristics. It was demonstrated that model calibration is always a trade-off between a parameterization that is general enough to reproduce multiple hydrograph characteristics and an accurate representation of specific aspects. By the selection of an objective function the modeler not only defines the calibration focus, but also implicitly makes assumptions about the statistical nature of data. Given the highly skewed distribution of discharge data and simulation errors, in this thesis the value of non-parametric criteria for model calibration was explored. To this end, a modification of the popular Kling-Gupta model efficiency towards a more non-parametric objective function was proposed, whereby results indicated the promising value of such non-parametric based model calibration criteria.

In the second part of this thesis, the value of single discharge observation for the prediction of discharge in otherwise ungauged catchments was investigated. In practice, single discharge observations could be collected during short field trips within a hydrological year. Such field trips were mimicked by strategically sampling observations from continuous discharge time series. A small number of discharge observations could already be informative for model calibration and therefore strongly improve discharge prediction compared to an ungauged situation. Although the optimal timing for making discharge measurements differed when aiming at the prediction

of hydrographs or flow-duration curves, good results could be achieved by measuring both the full range of a catchment's discharge magnitudes and its major runoff events. The value of such single discharge observations was furthermore tested for the prediction of discharge using regionalization. Thereby, a small number of discharge observations could effectively improve regionalization with attribute similarity and spatial proximity, especially in catchments with a distinct runoff regime or a pronounced high-flow period.

Keywords: hydrological modeling, HBV, model calibration, objective function, value of data, ecological streamflow characteristics, single discharge observations, sampling strategy, regionalization, large-sample data set, United States

Zusammenfassung

Abflussdaten sind eine wichtige Grundlage für Entscheidungen in der Wasserwirtschaft. Solche Daten fehlen jedoch in vielen Einzugsgebieten. In ungemessenen Einzugsgebieten werden daher oftmals Abflusszeitserien mithilfe von hydrologischen Modellen simuliert. Hydrologische Modelle beinhalten Parameter, welche die verschiedenen Speicher und Flüsse von Wasser in einem natürlichen Gebiet quantifizieren. Die Werte dieser Parameter können typischerweise nicht direkt gemessen werden, da die Parameter einerseits eine konzeptionelle Bedeutung haben und anderseits eine Diskrepanz zwischen Mess- und Modellskala besteht. Aus diesem Grund müssen Modellparameterwerte durch eine Kalibration geschätzt werden. Die Parameterwerte werden geschätzt, indem die Differenz zwischen simuliertem und beobachtetem Abfluss anhand eines Gütekriteriums minimiert wird. Der Bedarf an zuverlässigen Abflusssimulationen und die Herausforderung der Modellkalibration für Gebiete ohne Abflussmessungen bilden den Kern dieser Dissertation.

Im ersten Teil dieser Dissertation befasste ich mich mit dem Einfluss der Modelkalibration, und somit dem Einfluss des Gütekriteriums, auf die Simulation ökologisch relevanter Abflussmerkmale. Solche spezifische Aspekte des Abflussverhaltens bilden oftmals die Grundlage einer nachhaltigen Wassernutzung. Im Rahmen dieser Dissertation wurde gezeigt, dass mit Simulationen basierend auf herkömmlichen Gütekriterien nicht alle ökologisch relevanten Abflussmerkmale zufriedenstellend abgeschätzt werden können. Sobald jedoch die Abflussmerkmale direkt in die Kalibration miteinbezogen werden, verbessert sich deren Simulationsgenauigkeit deutlich. Zusammenfassend kann gesagt werden, dass die Modellkalibration ein Kompromiss ist zwischen dem Ziel der Simulation spezifischer Abflussmerkmale und dem Ziel eine möglichst grosse Vielfalt an Merkmalen mit hoher Genauigkeit zu Simulieren. Die Wahl des Gütekriteriums beeinflusst nicht nur den Fokus der Kalibration sondern impliziert auch statistische Annahmen über Abflussdaten. Da die statistische Verteilung von Abflussdaten und Modellfehlern deutlich von einer Normalverteilung abweichen, wurde in dieser Dissertation eine modifizierte Variante des weitverbreiteten Kling-Gupta Kriteriums getestet, welche nicht-parametrische Komponenten enthält. Die Resultate zeigten, dass nicht-parametrischer Kriterien eine sinnvolle Alternative zu herkömmlichen Kriterien darstellen.

Der Fokus im zweiten Teil dieser Dissertation lag auf dem Wert von vereinzelten nicht kontinuierlich gemessenen Abflussdaten für die Modellkalibration. Solch vereinzelte Abflussdaten könnten zum Beispiel durch kurze Feldbesuche in einem Gebiet ohne kontinuierliche Zeitserien erhoben werden. Die Resultate dieser Dissertation zeigten auf, dass bereits zwölf Messungen zu strategisch wichtigen Zeitpunkten einen hohen Informationsgehalt für die Modelkalibration enthalten. Der exakte Zeitpunkt für die informativste Messung ist je nach Simulationszweck, wie beispielsweise der Simulation von Hydrographen und Dauerkurven, verschieden. Dennoch können mit einer gezielten Messstrategie, die sowohl ein hohes Abflussereignis als auch die

charakteristische Abflussverteilung wiederspiegelt, mehrere Hydrographaspekte deutlich besser simuliert werden als wenn keine Abflusswerte zur Verfügung stehen. Solch strategisch gemessene Abflussdaten können zudem verwendet werden, um die Simulation von Abfluss mittels Regionalisierung zu verbessern. Ihr Wert ist dabei besonders hoch in Gebieten mit einem ausgeprägten Abflussregime oder einer ausgeprägten Periode mit hohem Abfluss.

Schlüsselwörter: Hydrologische Modellierung, HBV, Modellkalibration, Gütekriterium, Wert von Daten, ökologisch relevante Abflussmerkmale, vereinzelte nicht kontinuierliche Abflussmessungen, Messstrategie, Regionalisierung, 'large-sample' Datensätze, Vereinigte Staaten von Amerika

Papers and Author Contributions

List of papers

I. Vis, M., R. Knight, S. Pool, W. Wolfe, and J. Seibert (2015), Model calibration criteria for estimating ecological flow characteristics, *Water*, *7*(5), 2358-2381.

II. Pool, S., M. Vis, R. Knight, and J. Seibert (2017), Streamflow characteristics from modelled runoff time series – importance of calibration criteria selection, *Hydrology and Earth System Sciences*, *21*(11), 5443-5457.

III. Pool, S., M. Vis, and J. Seibert (in press), Evaluating model performance: Towards a non-parametric variant of the Kling-Gupta efficiency, *Hydrological Science Journal*

IV. Pool, S., D. Viviroli, and J. Seibert (2017), Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration?, *Journal of Hydrology*, 554, 613-622.

V. Pool, S., D. Viviroli, and J. Seibert, Value of a limited number of discharge observations for improving regionalization: A large sample study across the United States, *resubmitted after minor revisions to Water Resources Research*.

Author contributions

Paper I: Rodney Knight and Jan Seibert conceived the initial ideas for this study. Marc Vis performed the simulations. Together with Jan Seibert, Marc Vis, and Rodney Knight, I contributed to the analysis and interpretation of the results. All authors contributed to writing of the manuscript.

Paper II: I designed this study together with Marc Vis, Rodney Knight and Jan Seibert based on the collaborative work in Paper I. Marc Vis performed the runoff simulations. I analyzed the simulations, whereby the results of the analysis were discussed with all co-authors. The first draft of the manuscript was written by myself and further developed with contribution of all co-authors.

Paper III, IV, and V: I had the lead in designing, implementing, analyzing, and writing the manuscript of these three studies. The co-authors, Marc Vis, Daniel Viviroli, and Jan Seibert shaped the studies by extensively discussing results with me and commenting on the manuscripts. Catchment input data for the hydrological model were processed by Marc Vis. Marc Vis also performed the mathematical implementation of the (partly) non-parametric Kling-Gupta efficiency.

Table of Contents

Abbreviations

The following list of abbreviations consists of a selection of abbreviations that are occurring frequently in this thesis.

Generic abbreviations

CAMELS	Catchment Attributes and MEteorology for Large-sample Studies (data set)
FDC	Flow-duration curve
HBV	Hydrologiska Byråns Vattenbalansavdelning (runoff model)
PUB	Prediction in Ungauged Basins
SFC	Streamflow characteristic

HBV model parameters (see Table 3.2 for more details)

P_{BETA}	Shape coefficient
P_{CFMAX}	Degree-day factor
P_{CFR}	Refreezing coefficient
P_{CWH}	Water holding capacity
P_{FC}	Max. soil moisture storage
P_{K0}	Recession coefficient of fast response
P_{K1}	Recession coefficient of intermediate response
P_{K2}	Recession coefficient of baseflow
P_{LP}	Threshold for reduction of evaporation
P_{MAXBAS}	Length of weighting function
P_{PERC}	Percolation from shallow to deep groundwater box
P_{SFC}	Snowfall correction factor
P_{TT}	Threshold temperature
P_{UZL}	Max. storage in shallow groundwater box

Sampling strategies (see Fig. 3.4 for more details)

$C_{Max\_Min\_Wetness}$	Samples of max. and min. discharge after different wetness conditions
C_{Max_Rec1}	Samples of peak discharge and event recession
C_{Max_Rec2}	Samples of peak discharge and event recession
$C_{Max\_Rec\_DOM}$	Samples of peak discharge, event recession, and at a fixed day of a month
$C_{Max\_Snowmelt}$	Samples of peak discharge and event recession during the snowmelt season
$I_{Max\_Min\_DOM}$	Samples of max. and min. discharge, and at a fixed day of a month
$I_{Mean\_Seasonal}$	Samples of seasonally biweekly mean discharge
$I_{Quantile}$	Samples of twelve discharge quantiles

Sampling strategies (cont.)

S_{DOM}	Samples at a fixed day of a month
S_{Max}	Samples of the monthly max. discharge
S_{Max_Min}	Samples of bimonthly min. and max. discharge
S_{Mean}	Samples of the monthly mean discharge
S_{Min}	Samples of the monthly min. discharge

Statistical metrics (see Table 3.4 or Fig. 3.3 for more details)

C_1	Efficiency metric consisting of R_{eff}, R_{eff_logQ}, and R_{VE}
C_2	Efficiency metric consisting of R_{eff}, R_{MARE}, $Spearman$, and R_{VE}
C_3	Efficiency metric consisting of $Spearman$ and R_{VE}
$Pearson$	Pearson correlation
R_{α}	Non-parametric efficiency metric for flow variability
R_{α_r}	Non-parametric efficiency metric for flow variability and flow dynamics
R_{β}	Efficiency metric for bias in flow volume
R_{β_α}	Efficiency metric for bias in flow volume and (non-parametric) flow variability
R_{β_r}	Efficiency metric for bias in flow volume and (non-parametric) flow dynamics
R_{eff}	Nash-Sutcliffe efficiency
R_{eff_logQ}	Efficiency for low flows
R_{eff_peak}	Efficiency for peak flows
R_{FDC}	Mean absolute error of the FDC at 99 evaluation points
R_{KG}	Kling-Gupta efficiency
R_{KG_α}	Kling-Gupta efficiency with a non-parametric variability component
R_{KG_r}	Kling-Gupta efficiency with a non-parametric correlation component
$R_{Lindström}$	Lindström measure
R_{MARE}	MARE measure
R_{NP}	Modified Kling-Gupta efficiency towards a non-parametric metric
R_r	Non-parametric efficiency metric for flow dynamics
R_{VE}	Volume error
$Spearman$	Spearman rank correlation

Streamflow characteristics (see Table 3.3 for more details)

DH13	Average 30-day max. runoff
DH16	Variability of high-flow pulse duration
E85	Lowest 15% of daily runoff
FH6	Frequency of moderate floods
FH7	Frequency of larger floods
FL2	Variability of low-flow pulse count
MA26	Variability of March runoff
MA41	Mean annual runoff
MH10	Max. October runoff
ML20	Baseflow
RA7	Rate of runoff recession
TA1	Stability of runoff
TL1	Timing of annual min. runoff

Streamflow characteristics-based metrics (see Table 3.5 or Chpt. 3.3.2 for more details)

B_{FDC}	Percent bias in the slope of the mid-flow segment of the FDC
B_{hf}	Percent bias in the high-flow segment of the FDC
B_{lf}	Percent bias in the low-flow segment of the FDC
B_{rr}	Percent bias in runoff ratio
B_t	Difference in watershed lag time
I_{Multi}	Efficiency for selected streamflow characteristics
I_{Multi_Reff}	Streamflow characteristics and model efficiency R_{eff}
I_{Single}	Efficiency for each individual streamflow characteristic
I_{Single_Reff}	Streamflow characteristic and model efficiency R_{eff}

CHAPTER 1

Introduction

1.1 Need for hydrological modeling

Hydrological models are often needed to support decision making in water management, when the availability of discharge observations in space and time is limited (*Beven*, 2012). Common applications for modeled discharge include the extension of streamflow records, the prediction of discharge in ungauged catchments, or the evaluation of the effect of climate change or land use change on various hydrograph characteristics (*Klemes*, 1986).

Hydrological models consist of a number of storages and fluxes of water, which represent the hydrological functioning of a catchment. These storages and fluxes are quantified by parameters. Typically, model parameter values cannot be directly derived from measurements in the field but have to be estimated by calibration. The need for calibration is mainly given by two reasons. First, model parameters are a conceptual approximation to reality and generally do not have a direct physical meaning. Second, model parameters with physical meaning are measured at the point scale but are representative for a much larger scale in the model (*Beven*, 2012). Estimation of parameter values by calibration is usually done by minimizing the difference between observed and simulated discharge using an error metric, also called objective function. Model calibration involves the decision on an objective function and relies on observed data records. The effect of the objective function on discharge simulations and approaches to predict discharge in data scarce catchments are outlined in the sections below.

1.2 Evaluating model performance

Model calibration traditionally relies on discharge information. However, depending on the model structure or data availability in a catchment, additional information, such as groundwater dynamics (*Pfannerstill et al.*, 2017; *Seibert and McDonnell*, 2015; *Juston et al.*, 2009), soil moisture state (*Shafii et al.*, 2017; *Hughes et al.*, 2014), or snow water equivalent (*Hingray et al.*, 2010) can be used for model calibration. Such additional data has also proved to be valuable in the form of soft data, i.e., qualitative knowledge of an experimentalist gained during field campaigns (*Seibert and McDonnell*, 2002). Yet, discharge is the most abundant type of data since it can be measured with limited effort and is an integrated measure of runoff response and storage dynamics in a catchment. Discharge is likely the first water balance component to be measured in case of ungauged or poorly gauged catchments. The following sections therefore focus on the use of discharge data for model evaluation.

1.2.1 Statistical metrics

Model calibration is historically based on statistical metrics like mean squared error, coefficient of determination, or volume error. Exhaustive lists of statistical metrics, their advantages and limitations are provided by e.g. *Krause et al.* (2005) or *Efstratiadis and Koutsoyiannis* (2010). Among statistical metrics, the Nash-Sutcliffe efficiency (*Nash and Sutcliffe*, 1970) is one of the most widely used ones to communicate model performance in hydrology. It is a dimensionless metric based on the ratio of the mean squared error between observed and simulated discharge and the variance in observed discharge. The use of the variance for normalizing the error term has been questioned given that the observed variance strongly depends on the runoff regime of a catchment (*Schaefli and Gupta*, 2007; *Krause et al.*, 2005). *Murphy* (1988) and *Gupta et al.* (2009) have shown that the mean squared error can be decomposed into the error in volume, variability and dynamics (i.e., correlation). Model calibration on the mean squared error has two disadvantages, which are the underestimation of discharge variability and the variable importance of discharge volume in calibration as a function of a catchment's discharge variability (*Gupta et al.*, 2009). As a consequence, *Gupta et al.* (2009) introduced a new efficiency measure (the Kling-Gupta efficiency) that consists of an improved combination of volume, variability, and timing errors. The three error components of the Kling-Gupta efficiency are expressed in terms of the bias in mean discharge, bias in discharge variability, and Pearson correlation coefficient. The Kling-Gupta efficiency is implicitly based on the assumptions of linearity and normality, and the absence of outliers. Since discharge observations and model errors are known to be highly skewed and usually contain outliers it has been proposed to represent discharge dynamics by the Spearman rank correlation instead of the Pearson correlation (*Legates and McCabe*, 1999). However, due to the use of ranks instead of absolute values, Spearman rank correlation is affected by a loss of information and there is not much research yet on how this influences the performance

of hydrological models. Both, Spearman and Pearson correlation are insensitive to additive and proportional volume errors, which is why correlation metrics should preferably be combined with volume error metrics for calibration (*Legates and McCabe*, 1999; *Krause et al.*, 2005).

1.2.2 Hydrological signatures

Hydrological signatures (also called streamflow characteristics) are indices derived from discharge time series. As opposed to statistical metrics, hydrological signatures are usually applied to describe specific aspects of the runoff response that have a physical meaning (*McMillan et al.*, 2017). Signatures used to evaluate model performance have been selected based on many different criteria. For example, *Viglione et al.* (2013) evaluated model performance using signatures that are of interest in various fields of water management. Selected signatures therefore included mean annual discharge, range of Pardé coefficient, and low-flow and flood metrics. *Yilmaz et al.* (2008) used signatures representing primary functions of a watershed for a process-based diagnostic model evaluation. The primary functions, i.e., water balance, vertical redistribution of water and runoff timing, were described by the signatures runoff ratio, slope of the flow-duration curve (*Vogel and Fennessey*, 1995), and lag-time. And finally, *Shrestha et al.* (2014) assessed discharge simulations in terms of ecologically relevant signatures (commonly referred to as ecologically relevant streamflow characteristics, SFC), which define the structure and functioning of aquatic and riparian biodiversity (*Poff et al.*, 1997; *Richter et al.*, 1996). Ecologically relevant signatures are usually site-specific, but include for example magnitude of annual flood, timing of low flows or rate of hydrograph recession. The selection of hydrological signatures is, however, as pointed out by *McMillan et al.* (2017), also often based on rather subjective decisions. They therefore proposed a selection guideline, which can assist hydrologists in choosing signatures by evaluating the five criteria of identifiability, robustness, consistency, representativeness, and discriminatory power.

Since hydrological signatures allow evaluating model performance for specific purposes, they are becoming more and more prominent in the hydrological modeling community. Signatures have been utilized for both model calibration and model validation. In model calibration, they were included with the intention to guide the selection of parameter values in a more meaningful way (*Yilmaz et al.*, 2008). Especially metrics derived from the flow-duration curve have been widely implemented as objective functions, whereby overall hydrograph simulations could be improved compared to simulations based on purely statistical objective functions (*Yilmaz et al.*, 2008; *Westerberg et al.*, 2011; *Pfannerstill et al.*, 2014). Also more specific signatures, such as the timing of spring snowmelt or the variability in low flow pulse count, have been used as objective functions. However, these signature-tailored calibrations resulted in accurate predictions of the respective signatures at the expense of other hydrograph aspects (*Hingray et al.*, 2010; *Olsen et al.*, 2013; *Kiesel et al.*, 2017). Similarly, multiple studies have shown that hydrological signatures calculated from simulated time series are often inadequately modeled for calibration approaches

on traditional statistical metrics (*Shrestha et al.*, 2014; *Ryo et al.*, 2015; *Murphy et al.*, 2013). It therefore remains a challenge to calibrate models in a way that simulated discharge time series represent a variety of signatures at high level of accuracy.

1.3 Effect of data record length on model performance

Modeling studies usually make use of long discharge time series to estimate model parameter values. However, even in regions considered as densely monitored many catchments are ungauged (i.e., have no or only limited discharge data), while at the same time the number of operated discharge stations is worldwide decreasing (*GRDC*, 2018). As part of the IAHS decade on prediction in ungauged basins (PUB) the question on the value of discharge data in constraining and reducing predictive uncertainty in ungauged basins got more attention in the hydrological community (*Sivapalan et al.*, 2003).

The first studies that pursued the question on the minimum length of discharge data needed for model calibration focused on continuous time series. Between two and eight years of discharge observations were reported as a minimum requirement for acceptable model simulations (*Harlin*, 1991; *Yapo et al.*, 1996; *Xia*, 2004; *Vrugt et al.*, 2006; *Merz et al.*, 2009). Although there is a general agreement that the identifiability of parameter values improves with increasing discharge record length (*Brath et al.*, 2004; *Perrin et al.*, 2007; *Rode et al.*, 2007; *Singh and Bárdossy*, 2012; *Tada and Beven*, 2012; *Vrugt et al.*, 2006; *Merz et al.*, 2009), relatively short or discontinuous discharge records can be of comparable value as long continuous time series. For example, *Brath et al.* (2004), *Melsen et al.* (2014), or *Sun et al.* (2017) reported that short continuous time series of a single season or a few months provide as much information for model calibration as records of a few years. The redundancy of information in discharge time series is also reflected in the fact that a comparable model performance can be achieved by a time series of multiple years and a random subset thereof (*Perrin et al.*, 2007; *Kim and Kaluarachchi*, 2009). Given the surprisingly high information content of randomly selected discharge observation, it can be hypothesized that a strategic selection of observations could further lower the amount of data needed for model calibration.

Using a set of simple strategies to extract observation from observed time series, *Seibert and Beven* (2009) and *Seibert and McDonnell* (2015) concluded that observations during high flows and discharge events are more informative for model calibration than mean and low-flow observations. Similarly, results of *Singh and Bárdossy* (2012) indicate the value of unusual events, such as high or low-flow periods or periods with strong discharge dynamics, for constraining model parameter values. Overall, limited data can be of comparable value as continuous time series as long as the available observations characterize dominant runoff processes and discharge variability of a catchment (*Harlin*, 1991; *Vrugt et al.*, 2006; *Konz and Seibert*, 2010; *Singh and Bárdossy*, 2012). Model calibration on short or discontinuous discharge observations is most reliable in humid

catchments (*Perrin et al.*, 2007; *Sun et al.*, 2017), whereby as few as 16 to 32 observations can already contain enough information for acceptable model simulations (*Seibert and Beven*, 2009). Most studies so far focused on the value of data for the simulation of hydrographs. From the perspective of water management in ungauged catchments it is furthermore of interest how the value of discharge observations varies between different hydroclimates and different hydrological signatures.

1.4 Prediction in ungauged catchments

In the absence of any discharge data, model parameter values cannot be estimated by common calibration approaches as outlined in the previous sections. To estimate parameter values in data scarce regions, a number of regionalization approaches have been proposed (for reviews see e.g. *He et al.*, 2011; *Parajka et al.*, 2013; *Razavi et al.*, 2013). The concept of regionalization relies on the idea of transferring hydrological information from gauged to ungauged catchments (*Blöschl and Sivapalan*, 1995). Regionalization approaches can be assigned to two major categories (*Parajka et al.*, 2013) that either transfer entire parameter sets within hydrologically similar regions (*Burn*, 1990) or that relate individual model parameters to catchment attributes (e.g., *Seibert*, 1999; *Kokkonen et al.*, 2003; *Merz and Blöschl*, 2004). Although the best regionalization method is likely site-specific (*He et al.*, 2011; *Razavi et al.*, 2013), a transfer of entire parameter sets is favored, because it accounts for parameter dependence (*Bárdossy*, 2007; *Buytaert and Beven*, 2009; *Kokkonen et al.*, 2003; *McIntyre et al.*, 2005).

Spatial proximity and attribute similarity are among the most widely applied regionalization approaches that use entire parameter sets from one or multiple donor catchments(s). Spatial proximity is based on Tobler‘s first law of geography that "everything is related to everything else, but near things are more related than distant things" (*Tobler*, 1970, p.236). It assumes that hydrological response varies smoothly in space (*Parajka et al.*, 2013), which allows to use the geographical distance between two catchments as an indicator for their similarity. The distance between catchment outlets (*Parajka et al.*, 2013), catchment centroids (*Arsenault and Brissette*, 2014; *Oudin et al.*, 2008; *Samuel et al.*, 2011), and a combination thereof (*Lebecherel et al.*, 2016) have been suggested as metrics for spatial proximity. Regionalization with attribute similarity builds on the concept that runoff response is governed by a combination of catchment attributes (*Burn*, 1990; *Burn and Boorman*, 1992). The approach therefore uses attributes to define hydrologically similar regions from which donor catchments can be extracted. Selected catchment attributes are usually descriptors of topographical aspects, land cover, climatic conditions, or soil type and geology (*Arsenault and Brissette*, 2014; *Merz and Blöschl*, 2004; *Oudin et al.*, 2008; *Zhang and Chiew*, 2009).

Regionalization can be associated with considerable uncertainties (*Hrachowitz et al.*, 2013). To reduce prediction uncertainty *Viviroli and Seibert* (2015) and *Rojas-Serna et al.* (2016) assumed

that a limited number of discharge observations could be collected in the ungauged catchment, which can be used to further constrain the values of regionalized model parameters. Their modeling results for catchments in Switzerland and France indicate an improved prediction efficiency when informing regionalization with a few observations as opposed to a classical regionalization. The value of such discontinuous but strategically sampled observations is most pronounced for catchments with a strong seasonal discharge regime (*Viviroli and Seibert*, 2015). This ultimately raises the question about the value of data for the prediction of discharge in different hydroclimates. Implementing the idea of an informed regionalization in practice will also require knowledge about the number of observations and their value when taken in years with different weather and discharge conditions.

CHAPTER 2

Scope of the Thesis

Many decisions in water management rely on accurate simulations of continuous discharge time series. This thesis makes a contribution to this challenge by evaluating the value of data in hydrological modeling. The value of data was evaluated in terms of the value of objective functions for the prediction of ecologically relevant streamflow characteristics (SFCs) and the value of a limited number of discharge measurements for model calibration and regionalization. More specifically, the following research questions were addressed within this thesis:

1. **How well are ecologically relevant SFCs estimated by simulations based on traditional model calibration criteria?**

   Streamflow characteristics are often calculated from simulated discharge time series that are based on model calibration with statistical metrics, such as the Nash-Sutcliffe efficiency or the volume error. In Paper I, it was evaluated if such traditional model calibration criteria preserve ecologically relevant SFCs commonly used in water management. To this end, 12 SFCs were estimated from simulations with seven statistical metrics for 27 catchments in the Tennessee River Basin, which has a very diverse freshwater ecosystem.

2. **Are SFC estimates improved by including them explicitly into model calibration?**

   Ecologically relevant SFCs cover a wide range of hydrograph characteristics including magnitude, frequency, duration, and timing of discharge. These rather specific characteristics are typically not all well simulated with model calibrations on commonly used statistical metrics. In Paper II, it was therefore hypothesized that including SFCs into

model calibration improves their estimate from simulated time series. This hypothesis was tested using the same study region and SFCs as in Paper I.

3. **Are non-parametric model calibration criteria a useful alternative to traditional criteria?**

   By selecting objective functions for model calibration one implicitly makes assumptions about the statistical nature of data. Since observed discharge time series and model simulation errors are known to be highly skewed, a modified formulation of the popular Kling-Gupta efficiency towards a non-parametric metric was proposed in Paper III. The three error components of the modified variant consisted of the mean discharge volume, the discharge variability expressed by the flow-duration curve and discharge dynamics described by the Spearman rank correlation. The proposed calibration criteria was evaluated by comparing simulations of various hydrograph aspects for 100 catchments spread across the contiguous United States.

4. **Which discharge measurements are most informative for model calibration in almost ungauged catchments?**

   The value of single discharge observations was addressed in Paper IV by assuming that a hydrologist gets the opportunity to take 12 discharge measurements within a hydrological year in an otherwise ungauged catchment. To mimic such recurring field visits, discharge measurements were strategically selected from observed time series. A total of 13 sampling strategies were defined to test the optimal timing for making measurements. The strategies were tested on 20 catchments with different degree of snow influence along the eastern United States.

5. **Does a limited number of discharge measurements improve regionalization?**

   Based on the knowledge gained in Paper IV, it was further tested if strategically taken discharge measurements provided valuable information for improving classical regionalization approaches (Paper V). The approach was tested in a leave-one out cross validation scheme on 579 catchments across the contiguous United States. The information value of discharge samples was compared between catchments, between sampling years, and for a varying number of measurements.

CHAPTER 3

Data and Methods

3.1 Study sites

3.1.1 Data set of the Tennessee River Basin

The studies about the prediction of SFCs (Papers I and II) were based on a set of 27 catchments in the Tennessee River Basin (Fig. 3.1a). The Tennessee River Basin has a very diverse freshwater ecosystem and has been exposed to drastic landscape changes due to urbanization, agriculture, and deforestation (*Abell et al.*, 2000). Modeling ecologically relevant SFCs in the Tennessee River Basin is of major interest since it allows to relate modeled changes in the runoff regime to potential changes in the ecosystem. The catchments selected for this thesis have all limited human influence. They are characterized by humid climatic conditions, mild precipitation seasonality, and negligible snowfall. Consequently, discharge is mostly precipitation driven with highest volumes in winter when soils are frozen and evaporation is low. For all catchments, continuous discharge (*U.S. Geological Survey*, 2014a), temperature and precipitation (*U.S. Department of Commerce*, 2007), as well as potential evaporation (*Rotstayn et al.*, 2006) data is available for a common time period of 28 years. Two catchments were discarded from the analysis of Paper II due to several weeks with missing discharge values, which was discovered during the analysis of the model simulations of Paper II.

3.1.2 Large-sample data set of the United States

The studies presented in Papers III to V were based on a large-sample data set of the United States. The data set was compiled by *Newman et al.* (2015) and consists of over 600 catchments with minimal human disturbances and 30 years of continuous daily discharge and meteorological time series. Discharge data were obtained from the *U.S. Geological Survey* (2014a), whereby less than 10 % of discharge values were missing for most of the catchments. The area averaged meteorological data was generated from the gridded (1 km by 1 km) Daymet data set (*Thornton et al.*, 2014). Since potential evaporation is not provided by the Daymet data set, it was calculated from other Daymet variables using the Priestley-Taylor equation (*Priestley and Taylor*, 1972). More recently, the *Newman et al.* (2015) data set was extended with a suite of catchment descriptors and is since then available as the CAMELS data set (*Addor et al.*, 2017). Attributes in the CAMELS data set include location and topography attributes, climate indices, soil characteristic, and vegetation characteristics. To complement the attributes provided by CAMELS, lake and wetland percentage were extracted for each catchment from a global data set of *Lehner and Döll* (2004). Additionally, SRTM data from *Jarvis et al.* (2008) was used to define elevation bands of 200 m for each catchment. Table 3.1 provides a summary of selected catchment attributes for the CAMELS data set.

Different subsets of the CAMELS data set were used within this thesis. Paper III was based on a stratified random subset of 100 catchments that covered the wide variety of hydroclimates in the Unites States. In Paper IV, twenty catchments from northeastern to southeastern United States were used, where changing snow conditions were the major climatic difference. Paper V was based on 579 catchments (Fig. 3.1b) for which the runoff model reproduced discharge at an acceptable level.

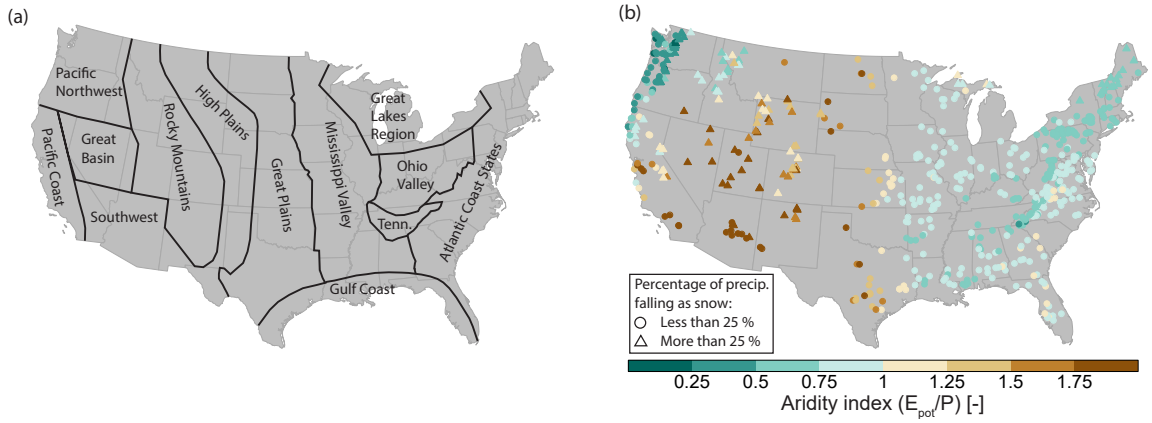


Figure 3.1: a) Geographical regions in the United States (adapted from *NOAA*, 2018). The Tennessee River Basin is labeled as 'Tenn.'. b) Locations of the 579 study catchments from the CAMELS data set (*Addor et al.*, 2017). Colors indicate the aridity index and the marker shape denotes the percentage of precipitation falling as snow.

Table 3.1: Statistics of catchment attributes of the 579 study catchments from the CAMELS data set (*Addor et al.*, 2017).

Catchment attribute	**5th quantile**	**Median**	**95th quantile**
Area [km^2]	22	301	2432
Aridity index[a] [-]	0.33	0.83	1.94
Precipitation seasonality[b] [-]	-1.13	0.06	0.65
Precipitation falling as snow [%]	0	9	69
Forested area [%]	2	86	100
Wetland area [%]	0	0	96
Clay content in soils [%]	6	19	36

Note: [a]Aridity index equals the ratio of sum of potential evaporation and sum of precipitation; [b]Precipitation seasonality is negative for catchments with winter precipitation, zero for catchments without precipitation seasonality, and positive for catchments with summer precipitation (for calculation see *Addor et al.*, 2017).

3.2 Hydrological model

The HBV (Hydrologiska Byråns Vattenbalansavdelning) runoff model (*Bergström*, 1976; *Lindström et al.*, 1997) was developed at the Swedish Meteorological and Hydrological Institute (SMHI) in the 1970s. Its structure, variables and parameters are based on physical considerations, such as the role of groundwater for runoff generation in northern latitudes, the empirical relationship between soil wetness and groundwater recharge, or the observation that hydrological conductivity decreases with soil depth. HBV was built under the premise of parsimony, i.e., with the aim of keeping the number of free parameters to a minimum. Thanks to it simplicity, the model has been applied in over 50 countries and exists in various variants. (*Bergström and Lindström*, 2015). The following section describes the structure of HBV, whereby text and equations are reproduced with some modification from *Seibert* (1999) and *Seibert and Vis* (2012).

HBV is a bucket-type runoff model that consists of four routines with a conceptual representation of snow pack dynamics, soil moisture variation, runoff response and discharge routing (Table 3.2 and Fig. 3.2). The model simulates continuous daily discharge time series using daily temperature and precipitation data and monthly potential evaporation data. Precipitation falls either as rain or as snow depending on the threshold temperature P_{TT}. Precipitation falling as snow is multiplied by a snowfall correction factor (P_{SFC}) to correct for systematic errors in precipitation measurements and evaporation from the snowpack. For daily temperatures ($T(t)$) above the threshold value P_{TT}, snowmelt M is calculated with the degree day method (Eq. 3.1), whereby melt rate is given by the degree day factor P_{CFMAX} and the temperature difference. Meltwater and rainfall can be retained in the snowpack until they exceed a certain fraction of water equivalent of the snow (P_{CWH}). Liquid water hold in the snowpack refreezes (R) at

temperatures below P_{TT} depending on the refreezing coefficient P_{CFR} (Eq. 3.2).

$$M = P_{CFMAX} \cdot (T(t) - P_{TT}) \tag{3.1}$$

$$R = P_{CFR} \cdot P_{CFMAX} \cdot (P_{TT} - T(t)) \tag{3.2}$$

Snowmelt and rainfall are input ($I(t)$) to the soil routine. The amount of water being stored in the soil or recharging the groundwater ($F(t)$) depends on wetness conditions, i.e. the ratio between the actual soil moisture content $S_{SOIL}(t)$ and its maximum value P_{FC} (Eq. 3.3). Soil moisture storage is furthermore depleted by actual evaporation (E_{ACT}). Actual evaporation equals potential evaporation (E_{POT}) if relative soil moisture is above a factor P_{LP}, while it linearly decreases for relative soil moisture contents below P_{LP} (Eq. 3.4).

$$\frac{F(t)}{I(t)} = \left(\frac{S_{SOIL}(t)}{P_{FC}}\right)^{P_{BETA}} \tag{3.3}$$

$$E_{ACT} = E_{POT} \cdot min\left(\frac{S_{SOIL}(t)}{P_{FC} \cdot P_{LP}}, 1\right) \tag{3.4}$$

Recharge from the soil routine contributes to the shallow groundwater storage S_{UZ} from which groundwater percolates to the deep storage S_{LZ} at a percolation rate P_{PERC}. Runoff from the two groundwater storages contributes to the peak, intermediate, and baseflow components of the hydrograph. The total daily groundwater runoff $Q_{GW}(t)$ is calculated as the sum of two or three linear outflow equation (K_0, K_1, K_2; Eq. 3.5), depending on whether S_{UZ} is above a threshold P_{UZL} that activates the peak runoff response. Finally, runoff from groundwater is transformed into the hydrograph $Q(t)$ at the catchment outlet by a triangular weighting function defined by the parameter P_{MAXBAS} (Eq. 3.6).

$$Q_{GW}(t) = P_{K2} \cdot S_{LZ} + P_{K1} \cdot S_{UZ} + P_{K0} \cdot max(S_{UZ} - P_{UZL}, 0) \tag{3.5}$$

$$Q(t) = \sum_{i=1}^{P_{MAXBAS}} c(i) \cdot Q_{GW}(t - i + 1)\ , \tag{3.6}$$

$$\text{where } c(i) = \int_{i-1}^{i} \frac{2}{P_{MAXBAS}} - \left|u - \frac{P_{MAXBAS}}{2}\right| \cdot \frac{4}{P_{MAXBAS}^2} \mathrm{d}u$$

In this thesis, HBV was applied in a semi-distributed form by disaggregating catchments into elevation bands of 200 m. Thereby, processes in the snow and soil routine are modeled for each elevation band separately, whereas the groundwater storage and the routing of the hydrograph are simulated in a lumped form. Temperature and precipitation input for the elevation bands were estimated by constant laps rates of 0.6°C per 100 m and 10 % per 100 m, respectively. Potential evaporation was assumed to be homogeneous across all elevation bands. The model version used in this thesis is HBV-light (*Seibert and Vis*, 2012).

Table 3.2: Specification of HBV model parameters (adapted from *Seibert and Vis*, 2012)

Parameter	Meaning	Unit	Min. value	Max. value
Snow routine				
P_{TT}	Threshold temperature	°C	-2	2.5
P_{SFC}	Snowfall correction factor	-	0.5	1.2
P_{CFMAX}	Degree-day factor	$mm°C^{-1}d^{-1}$	0.5	10
P_{CFR}	Refreezing coefficient	-	0	0.1
P_{CWH}	Water holding capacity	-	0	0.2
Soil routine				
P_{FC}	Max. soil moisture storage	mm	100	550
P_{BETA}	Shape coefficient	-	1	5
P_{LP}	Threshold for reduction of evaporation	-	0.3	1
Groundwater routine				
P_{UZL}	Max. storage in shallow groundwater box	mm	0	70
P_{PERC}	Percolation from shallow to deep groundwater box	mmd^{-1}	0	4
P_{K0}	Recession coefficient of fast response	d^{-1}	0.1	0.5
P_{K1}	Recession coefficient of intermediate response	d^{-1}	0.01	0.2
P_{K2}	Recession coefficient of baseflow	d^{-1}	0.00005	0.1
Routing routine				
P_{MAXBAS}	Length of weighting function	d	1	5

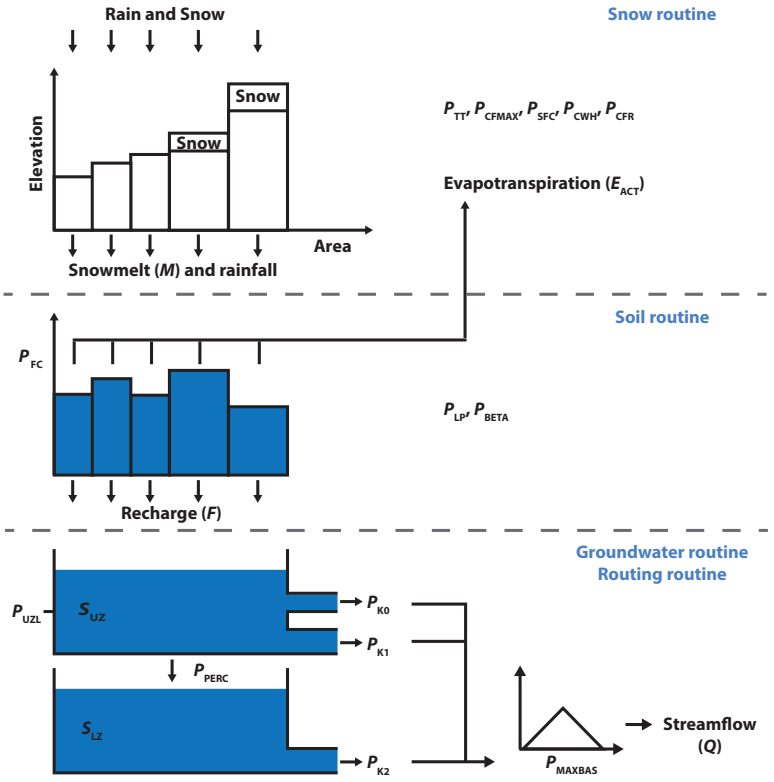


Figure 3.2: Structure, variables, and parameters of the HBV runoff model (adapted from *Uhlenbrook et al.*, 1999; *Bergström*, 1992).

3.3 Value of hydrograph characteristics for model calibration

The value of hydrograph characteristics (in terms of objective functions) for model calibration was evaluated in Papers I to III. While the focus in Paper I and Paper II was on the value of hydrograph characteristics for the prediction of ecologically relevant streamflow characteristics (Chpt. 3.3.1), the aim of Paper III was to evaluate the effect of using non-parametric metrics on various hydrograph aspects (Chpt. 3.3.2).

3.3.1 Prediction of ecologically relevant streamflow characteristics

Selection of streamflow characteristics

The SFCs assessed in this thesis (Table 3.3) were identified as important streamflow indicators for fish species diversity in the Tennessee River Basin (*Knight et al.*, 2008, 2014). Together they represent the five major flow regimes typically addressed in ecological studies: magnitude, ratio, frequency, variability, and date (e.g. *Olden and Poff*, 2003; *Arthington et al.*, 2006; *Caldwell et al.*, 2015). While model performance was evaluated for 12 SFCs in Paper I, an additional SFC representing low-flow conditions was selected for Paper II to have a balanced number of characteristics for high, mean and low-flow conditions. The SFCs were calculated from the observed and simulated discharge time series using the USGS EflowStats R-package (*U.S. Geological Survey*, 2014b).

Value of traditional calibration criteria

Continuous daily discharge time series were simulated for each catchment with HBV for two time periods of 14 years (1 October 1983 to 30 September 1996 and 1 October 1996 to 30 September 2009). Each time period served both as calibration and validation period, and the approximately three years preceding the simulation periods were used as warm-up period to establish reasonable state variables. In the calibration period, model parameters were optimized a 100 times within predefined parameter ranges (Table 3.2) using a genetic algorithm (*Seibert*, 2000) and commonly used calibration criteria. As commonly used calibration criteria, four statistical metrics that focus on different hydrograph aspects were selected. This selection included R_{eff}, R_{eff_logQ}, $R_{Lindström}$, and R_{MARE} (for formulas see Table 3.4). Additionally, three multi-objective functions were defined to calibrate HBV on different aspects simultaneously. Calibration metric C_1 was a combination of R_{eff}, R_{eff_logQ}, and R_{VE}, metric C_2 consisted of R_{eff}, R_{MARE}, Spearman rank correlation, and R_{VE}, and finally metric C_3 was composed of Spearman rank correlation and R_{VE}. The various single metrics in C_1, C_2, and C_3 were equally weighted.

For each of the seven calibration criteria, there were 100 simulations for each catchment during the calibration and validation period. SFCs were calculated from these 100 simulated hydrographs and were normalized by their observed value to obtain a percent error [%]. The final evaluation of the results was based on the median of all 100 percent error values.

Table 3.3: Description of streamflow characteristics used in Paper I and II (adapted from *Knight et al.*, 2014; *U.S. Geological Survey*, 2014b)

Streamflow characteristic	**Abbreviation**	**Further explanation**	**Flow condition**	**Unit**
Magnitude				
Mean annual runoff	MA41	Mean annual daily runoff	Mean flow	[mmd^{-1}]
Max. October runoff	MH10	Mean of October runoff maxima for each year	High flow	[mmd^{-1}]
Lowest 15% of daily runoff	E85	Daily mean runoff that is exceeded 85 % of the time for the period of record	Low flow	[mmd^{-1}]
Rate of runoff recession	RA7	Median change in log of runoff for days in which the change is negative across the period of record	Mean flow	[mmd^{-1}]
Ratio				
Average 30-day max. runoff	DH13	Mean annual max. of a 30-day moving average runoff divided by the median for the entire record	High flow	[-]
Baseflow	ML20	Ratio of total baseflow to total flow. Baseflow is the min. flow magnitude in a 5-day window if 90 % of that min. flow magnitude is less than the min. flow magnitude of the 5 day window before and after the considered window	Low flow	[-]
Stability of runoff	TA1	Measure of the constancy of a flow regime by dividing daily flows into predetermined flow classes. The 11 flow classes capture flow ranging from flow less than 0.1 times the logarithmic mean flow to flow more than 2.25 times the logarithmic mean flow	Mean flow	[-]
Frequency				
Frequency of moderate floods	FH6	Average number of high-flow events per year that are equal to or greater than three times the median annual flow for the period of record	High flow	[a^{-1}]
Frequency of larger floods	FH7	Average number of high-flow events per year that are equal to or greater than seven times the median annual flow for the period of record	High flow	[a^{-1}]
Variability				
Variability of March runoff	MA26	Standard deviation for March runoff over the period of record divided by the mean runoff for March over the period of record	Mean flow	[%]
Variability of high-flow pulse duration	DH16	Standard deviation for the yearly average high-flow pulse duration (daily flow greater than the 75th percentile) divided by the mean of the yearly average high-flow pulse duration multiplied by 100	High flow	[%]
Variability of low-flow pulse count	FL2	Standard deviation for the average number of yearly low-flow pulses (daily flow less than the 25th percentile) divided by the mean low-flow pulse counts multiplied by 100	Low flow	[%]
Date				
Timing of annual min. runoff	TL1	Julian date of annual min. flow occurrence	Low flow	[Julian day]

Table 3.4: Statistical metrics used for model evaluation.

Objective function	**Abbreviation**	**Definition**
Nash-Sutcliffe efficiency	R_{eff}	$1-\frac{\sum_{t=1}^{n}(Q_{obs}(t)-Q_{sim}(t))^2}{\sum_{t=1}^{n}(Q_{obs}(t)-\overline{Q_{obs}})^2}$
Efficiency for low flows	R_{eff_logQ}	$1-\frac{\sum_{t=1}^{n}(\ln Q_{obs}(t)-\ln Q_{sim}(t))^2}{\sum_{t=1}^{n}(\ln Q_{obs}(t)-\overline{\ln Q_{obs}})^2}$
Efficiency for peak flows	R_{eff_peak}	$1-\frac{\sum_{p=1}^{m}(Q_{obs}(p)-Q_{sim}(p))^2}{\sum_{p=1}^{m}(Q_{obs}(p)-\overline{Q_{obs}})^2}$
Volume error	R_{VE}	$1-\frac{\lvert\sum_{t=1}^{n}(Q_{obs}(t)-Q_{sim}(t))\rvert}{\sum_{t=1}^{n}(Q_{obs}(t))}$
Lindström measure	$R_{Lindström}$	$R_{eff}-0.1\frac{\lvert\sum_{t=1}^{n}(Q_{obs}(t)-Q_{sim}(t))\rvert}{\sum_{t=1}^{n}(Q_{obs}(t))}$
MARE measure	R_{MARE}	$1-\frac{1}{n}\sum_{t=1}^{n}\frac{\lvert Q_{obs}(t)-Q_{sim}(t)\rvert}{Q_{obs}(t)}$
Pearson correlation	$Pearson$	$\frac{\sum_{t=1}^{n}(Q_{obs}(t)-\overline{Q_{obs}})(Q_{sim}(t)-\overline{Q_{sim}})}{\sqrt{\sum_{t=1}^{n}(Q_{obs}(t)-\overline{Q_{obs}})^2}\sqrt{\sum_{t=1}^{n}(Q_{sim}(t)-\overline{Q_{sim}})^2}}$
Spearman rank correlation	$Spearman$	$\frac{\sum_{t=1}^{n}(R_{obs}(t)-\overline{R_{obs}})(R_{sim}(t)-\overline{R_{sim}})}{\sqrt{\sum_{t=1}^{n}(R_{obs}(t)-\overline{R_{obs}})^2}\sqrt{\sum_{t=1}^{n}(R_{sim}(t)-\overline{R_{sim}})^2}}$
Kling-Gupta efficiency	R_{KG}	$1-\sqrt{(\beta-1)^2+(\alpha_{KG}-1)^2+(r_p-1)^2}$
		$\beta=\frac{\overline{Q_{sim}}}{\overline{Q_{obs}}}$
		$\alpha_{KG}=\frac{\sigma_{Q_{sim}}}{\sigma_{Q_{obs}}}$
		$r_p=Pearson$
Kling-Gupta efficiency with non-parametric components	R_{NP}	$1-\sqrt{(\beta-1)^2+(\alpha_{NP}-1)^2+(r_s-1)^2}$
		$\alpha_{NP}=1-\frac{1}{2}\sum_{k=1}^{n}\left\lvert\frac{Q_{sim}(I(k))}{n\overline{Q_{sim}}}-\frac{Q_{obs}(J(k))}{n\overline{Q_{obs}}}\right\rvert$
		$r_s=Spearman$

Note: $Q_{obs}(t)$ and $Q_{sim}(t)$ are observed and simulated discharge at time step t; $R_{obs}(t)$ and $R_{sim}(t)$ are the ranks of $Q_{obs}(t)$ and $Q_{sim}(t)$; $Q_{obs}(p)$ is the observed peak flow value and $Q_{sim}(p)$ is the highest simulated discharge within three days of the observed peak (p); n is the length of the time series, m is the number of observed peaks within that time series, and σ is the standard deviation; $Q_{sim}(I(k))$ and $Q_{obs}(J(k))$ are the simulated and observed discharge with rank k.

Value of SFCs as calibration criteria

Discharge simulations in Paper II were based on the same modeling set-up as in Paper I with the difference that model calibration was explicitly targeted towards the 13 ecologically relevant SFCs of the Tennessee River Basin. First, new objective functions (Table 3.5) were defined that consisted of a single SFCs (I_{Single}). Each SFC was once used as objective function resulting in 13 I_{Single}. The individual SFCs (i.e. I_{Single}) were evaluated in terms of their robustness and their information value. Thereby, robustness was measured by how well a SFC was estimated when it was simulated using I_{Single}. A SFC was regarded as informative when it also yielded relatively good simulations for other SFCs. The four most robust and informative SFCs were then combined into a multi-objective function (I_{Multi}). Both I_{Single} and I_{Multi} were furthermore combined with the objective function R_{eff} to improve the overall shape of the simulated hydrograph, including magnitude and timing of events. Model performance for each of the described SFC-based objective function was evaluated using the median of the normalized SFC error for all 100 calibration runs. The normalized SFC error was calculated as the absolute simulation error between observed and simulated SFC divided by the range of possible SFC values in the respective catchment. The range of possible SFC values of each catchment was approximated by 10'000 Monte Carlo simulations, whereby the range was the difference between the 10th and the 90th quantile. Please note that simulation periods for Paper II were only 13 years (1 October 1984 to 30 September 1996 and 1 October 1997 to 30 September 2009) as opposed to 14 years in Paper I.

Table 3.5: Streamflow characteristics-based metrics used for model evaluation.

Objective function	**Abbreviation**	**Definition**
Efficiency for each individual SFC[1]	I_{Single}	$1-\frac{\lvert I_{obs}-I_{sim}\rvert}{I_{obs}}$
SFC and model efficiency	I_{Single_Reff}	$0.5(I_{Single}+R_{eff})$
Efficiency for the selected $SFCs$[2]	I_{Multi}	$\frac{1}{n}(I_{Single_1}+\ldots+I_{Single_n})$
SFCs and model efficiency	I_{Multi_Reff}	$\frac{n-1}{n}I_{Multi}+\frac{1}{n}R_{eff}$

Note: [1]For each of the 13 SFCs a specific I_{Single} exists; [2]I_{Multi} consists of the n most robust and informative SFCs.

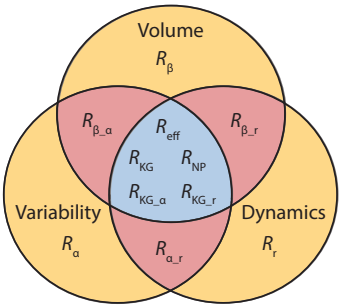
3.3.2 Towards a non-parametric variant of the Kling-Gupta efficiency

Objective functions

The Kling-Gupta efficiency (R_{KG}) considers three different types of model errors, namely the error in the mean, the variance and the dynamics. The three components are calculated as the bias in mean discharge, bias in the standard deviation of discharge, and the Pearson correlation between observed and simulated discharge time series (*Gupta et al.*, 2009). All three criteria are implicitly based on the assumption of data normality. To account for highly skewed distributions of discharge and simulation errors and to be less sensitive to outliers, the idea of a partly non-parametric formulation of R_{KG} was tested in Paper III. Similar as R_{KG}, the 'non-parametric' variant R_{NP} used the mean discharge as a measure of central tendency. However, discharge variability and dynamics were expressed by the flow-duration curve (FDC) and the Spearman rank correlation, respectively. The three components for mean discharge (β), discharge variability (α), and discharge dynamics (r) used in their parametric and non-parametric variants built the foundation of various one, two, and three-component objective functions (see Fig. 3.3). The three components β, α, r, as well as R_{KG} and R_{NP} are described in detail in Table 3.4.

Modeling approach

Each of the 11 objective functions (Fig. 3.3) was used to optimize HBV model parameters during a ten year time period (1 October 1990 to 30 September 1999). Model calibration was performed a 100 times within predefined parameter ranges (Table 3.2) using a genetic algorithm (*Seibert*, 2000). The 100 calibrated parameter sets were used to simulate discharge in an independent validation period (1 October 2000 to 30 September 2009). For both calibration and validation a two year warming-up period was used to ensure suitable initial values for the state variables. Discharge simulations in the validation time period were evaluated in two ways. First, it was evaluated how R_{KG} and R_{NP} affect simulations, and more specifically hydrograph uncertainty, during calibration. Hydrograph uncertainty was quantified by the difference between the 5th and 95th quantile of all 100 simulated hydrographs at each time step. The difference was then normalized by the observed discharge and evaluated for different discharge quantiles. Second, model performance was evaluated for i) the calibration metrics R_{KG}, R_{NP}, and R_{eff}, ii) three commonly used statistical metrics R_{eff_logQ}, R_{MARE}, and R_{eff_peak} (Table 3.4), and iii) for five hydrological signatures. The hydrological signatures were selected according to *Yilmaz et al.* (2008) and included the percent bias in runoff ratio (B_{rr}), the watershed lag time (B_t), the percent bias in the high-flow segment of the FDC (B_{hf}), the slope of the mid-flow segment of the FDC (B_{FDC}), and the low-flow segment of the FDC (B_{lf}). Model performance for the calibration metrics, the statistical metrics, and the signatures was evaluated using the median of all 100 calibration runs of each catchment.



	Volume (β)	Variability (α)	Dynamcis (r)
R_{eff}	Mean	Standard deviation	Pearson
R_{KG}	Mean	Standard deviation	Pearson
R_{KG_α}	Mean	Flow-duration curve	Pearson
R_{KG_r}	Mean	Standard deviation	Spearman
R_{NP}	Mean	Flow-duration curve	Spearman
R_{β_α}	Mean	Flow-duration curve	-
R_{β_r}	Mean	-	Spearman
R_{α_r}	-	Flow-duration curve	Spearman
R_β	Mean	-	-
R_α	-	Flow-duration curve	-
R_r	-	-	Spearman

Figure 3.3: Towards non-parametric objective functions. The three basic components describing discharge volume (β), variability (α), and dynamics (r) were used in their parametric and non-parametric variants and combined into a total of 11 one-, two- or three-component objective functions (colored in yellow, red and, blue respectively).

3.4 Gauging the ungauged catchment: Value of single discharge observations

In some cases a catchment lacks continuous discharge time series, but a limited number of discharge measurements could be taken during short field trips within a hydrological year. To mimic such field trips, a limited number of observations was strategically extracted from the observed discharge time series of each study catchment. In Paper IV, various sampling strategies to decide on when to measure discharge were defined and used to calibrate HBV (Chpt. 3.4.1). Based on the results of Paper IV, one sampling strategy was chosen to select discharge observations that were then used to inform classical regionalization approaches in Paper V (Chpt. 3.4.2).

3.4.1 Which discharge observations are most informative for model calibration?

Defining sampling strategies

A total of 13 strategies were defined considering both practical aspects and hydrological knowledge. All these strategies were restricted to 12 discharge observations within one hydrological year (Fig. 3.4). From the practical perspective of conducting recurring field trips, it was interesting to test rather simple strategies, such as making observations at a fixed day of a month or at event peaks. From a hydrological perspective, more complex strategies that sample dominant runoff processes or discharge variability could be promising. Therefore, strategies were defined to e.g. measure during the snowmelt season, multiple event recessions, wet and dry periods or different

discharge quantiles.

Modeling approach

The sampling strategies were used to retrieve 12 discharge observations from each of the 14 hydrological years (1983 to 1996) of each catchment. The 12 observations were then used to estimate HBV model parameters in a Monte Carlo approach. More specifically, 100'000 parameter sets were randomly generated and used to simulate discharge during all 14 sampling years. For each year, discharge simulations were evaluated by calculating R_{eff} and R_{eff_logQ} for the dates of the 12 observations. The 100 best parameter sets for each sampling year and strategy were then retained to additionally simulate discharge in an independent validation period (1 October 1997 to 30 September 2010).

From these 100 simulations, an ensemble mean hydrograph and FDC were calculated (Eq. 3.7). In both cases, the ensemble mean $\overline{Q}$ at each time step or at each of the 99 evaluation points t was calculated by equally weighting each (i) of the N simulations:

$$\overline{Q}(t) = \sum_{i=1}^{N} Q_i(t) W_i \tag{3.7}$$

Ensemble mean hydrograph performance was assessed in terms of R_{eff}, whereas the ensemble mean FDC was evaluated by the mean absolute relative error at 99 evaluation points of the FDC (R_{FDC}; see *Westerberg et al.*, 2011). The ensemble mean model performance of each sampling year and strategy (R_{ss}) was normalized by an upper (R_{ub}) and a lower benchmark (R_{lb}) according to *Seibert et al.* (2018):

$$R^* = \frac{R_{ss} - R_{lb}}{R_{ub} - R_{lb}} \tag{3.8}$$

The upper benchmark represented a well-informed model calibration with a continuous 14 year time period. In case of the hydrograph it was the ensemble mean of the 100 best parameter sets selected by R_{eff} or R_{eff_logQ} . The upper benchmark of the FDC was the ensemble mean of the 100 best parameter sets regarding R_{FDC}. The lower benchmark for both the hydrograph and the FDC was calculated as ensemble mean from 1000 random parameters sets and was an indication of model performance in the absence of any discharge data.

Additionally, we evaluated the effect of each sampling strategy on model parameter uncertainty. Parameter uncertainty was measured in terms of the range of the 100 parameter values (5th to 95th quantile) resulting from a particular sampling strategy. The observed range of parameter values was divided by the possible range of parameter values before calibration (Table 3.2) to enable a comparison across parameters.

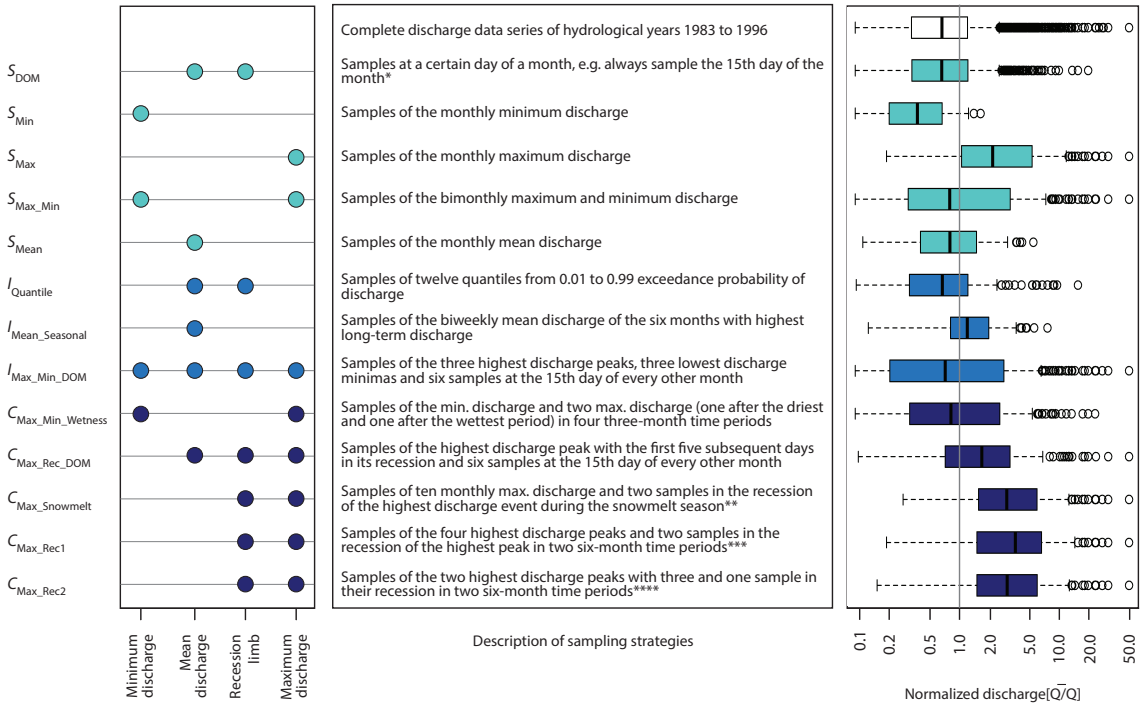


Figure 3.4: Definition of the 13 sampling strategies used to select single discharge observations (samples) for model calibration. Each sampling strategy consisted of 12 discharge samples. From left to right: abbreviation of sampling strategies, conceptual idea of discharge represented by strategies, description of strategies, and normalized discharge magnitudes sampled with the strategies (normalized discharge corresponds to the sampled discharge Q divided by the mean catchment discharge $\overline{Q}$ of a selected study catchment). *S_{DOM}: the strategy was tested with samples at the 1st, 5th, 10th, 15th, 20th and 25th day of the month, whereby the mean was calculated over the performance of all these six versions. **$C_{Max\_Snowmelt}$: maximum discharge of the ten months with highest long-term discharge and recession samples taken at 80 % and 60 % of highest discharge peak in the snowmelt season (February to May). ***C_{Max_Rec1}: recession samples taken at 80 % and 40 % of highest discharge peak. ****C_{Max_Rec2}: recession samples taken at 80 %, 60 % and 40 % of highest discharge peak and 80 % of second highest discharge peak.

3.4.2 Informing regionalization with a limited number of discharge observations

Classical regionalization

Regionalization was based on five donor catchments that provided their entire parameter sets to the ungauged catchment. Donor catchments were selected based on two commonly used regionalization approaches. First, spatial proximity calculated as the Euclidean distance (*Burn*, 1990; *McIntyre et al.*, 2005) between catchment centroids was used to select the five geographically closest catchments. Second, donor catchments were chosen that are similar in terms of catchment attributes. Here, the Euclidean distance between seven selected attributes was used: catchment area (log-transformed values), aridity, precipitation seasonality, percentage of precipitation falling as snow, percentage of forested area, percentage of wetland area, and percentage of clay content in soils (for explanations see Table 3.1). The attributes were standardized (Eq. 3.9; *Milligan and Cooper*, 1988) before calculating the Euclidean distance:

$$Z = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{3.9}$$

Z is the standardized attribute and X is the original attribute value. Figure 3.5 gives an impression of the median distance between the ungauged catchment and its five donors when selected with spatial proximity or attribute similarity.

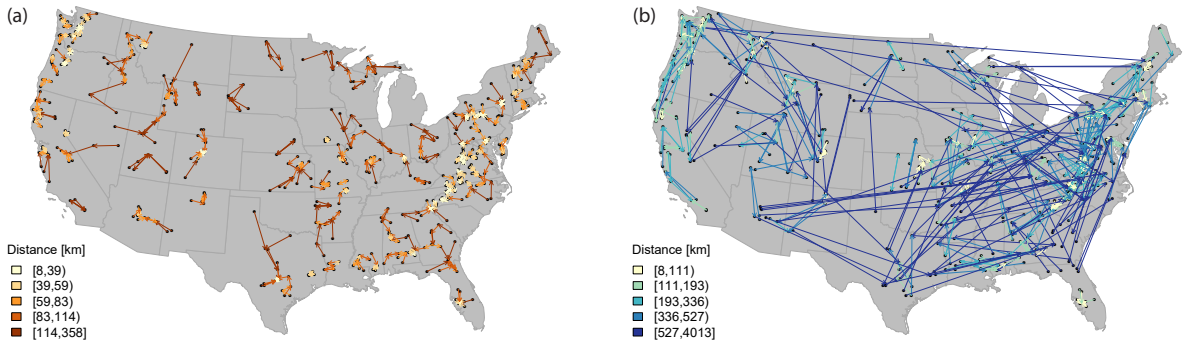


Figure 3.5: Location of the median donor catchment and the corresponding geographical distance in space (km) for the selection of donors with a) spatial proximity, and b) attribute similarity.

The calibration of HBV for each catchment, by optimizing R_{NP} a 100 times over a ten year calibration time period (1 October 1989 to 30 September 1999), built the foundation of the regionalization. Regionalization was conducted in a leave-one-out cross validation, where each catchment was once treated as ungauged at a time and received the parameter sets from its five donor catchments. Using the total of 500 donated parameter sets, discharge simulations were generated for the ungauged catchment in the calibration and the validation period (1 October 1999 to 30 September 2009). These simulations were combined by calculating an ensemble

mean hydrograph (Eq. 3.7) with equally weighted simulations. The ensemble mean hydrograph was evaluated using R_{NP}. The described regionalization approach with spatial proximity and attribute similarity will be referred to as *classical* regionalization.

Informing regionalization with observations

Based on the results of Paper IV, the concept of strategy $C_{Max\_Rec\_DOM}$ (Fig. 3.4) was used to extract 3, 6, 12, or 24 discharge observations from the observed time series of each catchment. The observations were extracted from each of the ten hydrological years of the calibration time period. They were used to evaluate the 500 discharge simulations from classical regionalization by calculating the root mean squared error (R_{RMSE}) between the observed and simulated discharge at the dates of the observations. The root mean squared error was then used to compute a weighted ensemble mean hydrograph (Eq. 3.7) in the validation time period, whereby the weight W_i of each parameter set i was calculated using Eq. 3.10 (where $R_{RMSE,max}$ is the highest R_{RMSE} among all parameter sets and N is the total number of j parameter sets ($j = 1,2,\ldots,N$)). The described regionalization approach that uses the information of single observations will be referred to as *informed* regionalization throughout this thesis.

$$W_i = \frac{\ln R_{RMSE,max} - \ln R_{RMSE,i}}{\sum_{j=1}^{N}(\ln R_{RMSE,max} - \ln R_{RMSE,j})} \tag{3.10}$$

The efficiency difference (ΔR_{NP}; Eq. 3.11) between the classical regionalization (R_{NP_CR}) with attribute similarity or spatial proximity and the informed regionalization (R_{NP_IR}) was used to evaluate the value of single observations between catchments, between sampling years, and for the varying number of observations.

$$\Delta R_{NP} = R_{NP\_IR} - R_{NP\_CR} \tag{3.11}$$

Catchments were compared by mapping ΔR_{NP} in space and by correlating ΔR_{NP} against catchment attributes (Spearman rank correlation). The effect of a sampling year on model performance was analyzed by calculating Spearman rank correlations between ΔR_{NP} and the hydrometerological conditions (e.g. sum of precipitation or peak discharge magnitude) of each sampling year. Finally, the effect of an increasing number of observations on the prediction efficiency was evaluated for sampling years with different information value. For a more detailed description of the model evaluation the reader is referred to Paper V of this thesis.

CHAPTER 4

Results

This chapter presents a summary of the main outcomes of this thesis. Fore more detailed results and additional figures and tables please see Papers I to V of this thesis.

4.1 Value of hydrograph characteristics for model calibration

4.1.1 Prediction of ecologically relevant streamflow characteristics

Value of traditional calibration criteria

Estimation accuracy of modeled SFCs varied considerably among the 27 study catchments. Therefore, taking the median over all catchments helped to focus the analysis on the general trends in the magnitude and the sign of the estimation accuracy. Table 4.1 presents these median estimation accuracies expressed in percent error for each of the seven objective functions. Since estimation accuracies of all SFCs were comparable in the two modeling time periods, only results for one modeling period are presented (calibration period 1983 to 1996).

Independent of the objective function used for model calibration, there was the tendency that SFCs representing mean (MA41, MA26, RA7, and TA1) and high-flow (MH10, DH13, DH16, FH6, and FH7) conditions were underestimated, whereas low-flow related SFCs (E85, FL2, TL1) were overestimated. On average, SFCs of mean-flow conditions were simulated with relatively high accuracy, i.e., percent errors between -2.8 % to -4.6 %, with the exception of RA7 that had an percent error of -41.1 %. Estimation accuracy for SFCs related to high flows were considerably lower than for average flows with percent errors ranging from -12.7 % to 32.5 %. Percent error for low-flow characteristics ranged between 4.1 % to 22.8 %.

Table 4.1: Estimation accuracy of simulated streamflow characteristics for model calibration with different objective functions. Estimation accuracy is expressed in terms of percent error [%]. Values indicate the median of the percent error for all 27 study catchments for the calibration period 1983 to 1996.

Objective function	**Mean flow**				**Low flow**			**High flow**				
	MA41	MA26	RA7	TA1	E85	FL2	TL1	MH10	DH13	DH16	FH6	FH7
R_{eff}	-2.5	9.8	-18.2	-10.8	9.8	17.5	4.2	-2.1	-14.7	-20.2	-12.0	-20.0
R_{eff_logQ}	-9.5	-7.3	-50.0	7.7	15.2	26.9	4.8	-20.0	-9.5	-10.0	-27.0	-37.5
$R_{Lindström}$	-0.6	9.1	-25.0	-15.2	19.1	16.8	3.7	-1.8	-18.1	-20.8	-12.0	-23.0
R_{MARE}	-18.9	-19.6	-57.1	25.0	-7.3	28.2	5.5	-44.0	-7.4	9.9	-41.4	-44.4
C_1	0.0	4.9	-50.0	-7.7	29.9	28.6	3.4	-4.8	-13.1	-19.7	-14.1	-19.0
C_2	-0.8	2.2	-42.9	0.0	13.2	17.7	4.0	-10.6	-7.5	-16.4	-18.2	-14.0
C_3	0.0	-28.1	-44.4	-18.9	24.1	23.6	3.4	-24.5	-18.9	-12.5	-37.6	-69.3
Average	-4.6	-4.2	-41.1	-2.8	14.9	22.8	4.1	-15.4	-12.7	-12.8	-23.2	-32.5

The objective function used to calibrate HBV did generally not change the sign of the percent error, but it strongly affected the magnitude of the percent error. Objective functions that emphasized high flows during calibration (i.e., included R_{eff}) resulted in highest prediction accuracy for high and mean-flow related SFCs. In contrast, the focus of the objective function was not necessarily a determinant for the estimation accuracy of low-flow related SFCs. Yet, no single best objective function could be observed that resulted in simulations accurately representing various hydrograph aspects. Also, the use of multi-objective functions instead of single-objective functions did not lead to better hydrograph simulations. Instead, it seemed to be more important that an objective function jointly evaluated the magnitude and the timing of discharge, which could be noted in the fact that the objective functions R_{eff_logQ}, R_{MARE}, and C_3 resulted in poor simulations for most SFCs.

Value of SFCs as calibration criteria

All 13 SFCs could be modeled with high accuracy during the calibration period if the SFC of interest was used as objective function (I_{Single}). Considering a SFC in calibration could clearly outperform model simulations based on the objective function R_{eff} for that particular SFC. The robustness of SFC-based model calibrations varied strongly between SFCs (Fig. 4.1a). SFCs related to physical catchment properties (e.g. RA7 or ML20) were the most robust ones among all tested SFCs, whereas SFCs that are subject to annual weather conditions (e.g. MH10 or TL1) were the least robust. The information value of SFCs, i.e., how informative I_{Single} was for other SFCs, was for many SFCs rather poor unless I_{Single} was combined with R_{eff} (I_{Single_Reff}).

Based on the concepts of robustness and information value, the SFCs MA41, RA7, ML20, and FH6 were selected as input for I_{Multi} and I_{Multi_Reff}. The four selected SFCs were not only among the most robust ones, but their combination ensured that each of 13 SFCs was relatively well simulated by model calibration with I_{Single} of either RA7, ML20, FH6 or MA41 (Fig. 4.1b). Together, the four selected SFCs provided information on the mean annual flow, the slope of the

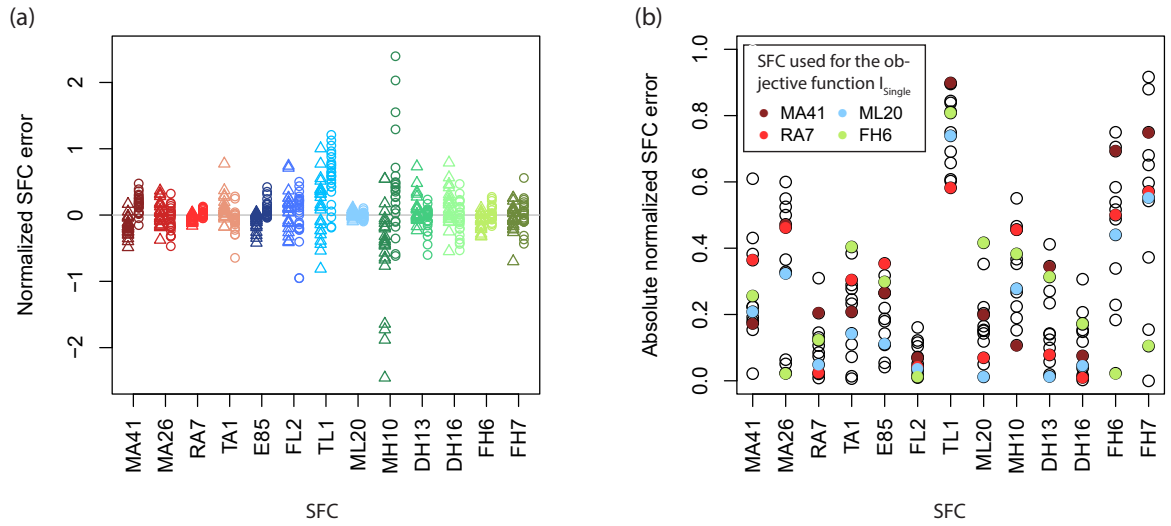


Figure 4.1: a) Robustness: normalized SFC errors in validation calculated from model calibrations with the objective function I_{Single} for the respective SFC. Values are shown for all 25 study catchments and both modeling time periods (triangles for period 1 (1984 - 1996) and circles for period 2 (1997 - 2009)). b) Information value: absolute normalized SFC errors in validation calculated from model calibrations with all 13 objective functions I_{Single}. Model performance values correspond to the median of the 25 study catchments and the mean of both modeling time periods. Each open circle represents one of the 13 SFC used for I_{Single}. The colored circles refer to the final selection of SFCs for the objective function I_{Multi}.

recessions, low-flow magnitudes, and the number of high flows. Results indicated that model calibrations with I_{Multi} or I_{Multi_Reff} led to comparable SFC estimates as model calibrations with R_{eff}, especially for SFCs not explicitly included in the objective function.

As previously observed in Paper I, HBV tended to underestimate mean and high flows, whereas low flows were generally overestimated (Fig. 4.2). The general trend of over- and underestimation could be considerably different for different objective functions or simulation time periods. Unlike the observations made in Paper I, there was no evidence that the magnitude of over- and underestimation was dependent on flow magnitudes.

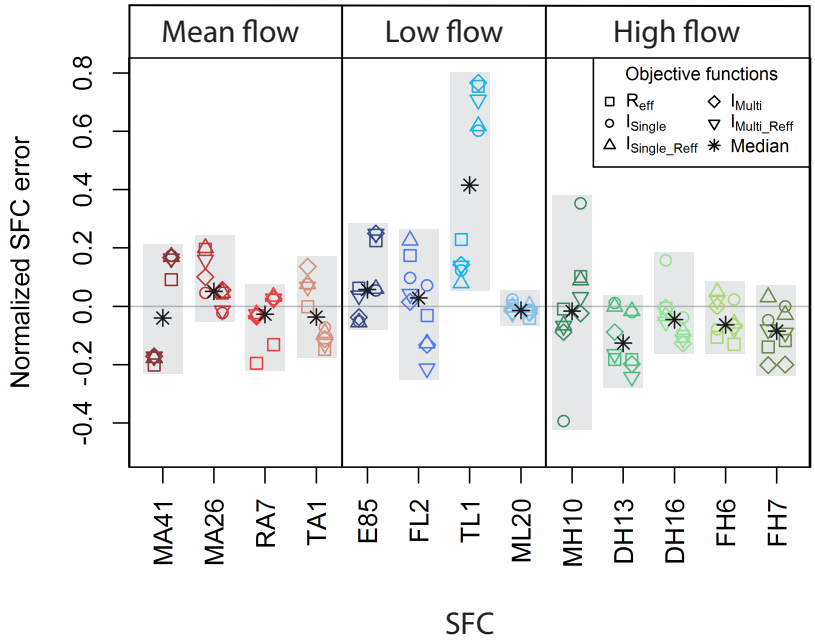


Figure 4.2: Normalized SFC errors in validation depending on the objective function used in calibration. Model performance values correspond to the median of the 25 study catchments and are shown for both modeling time periods (period 1 (1984 - 1996) on the left side and period 2 (1997 - 2009) on the right side).

4.1.2 Towards a non-parametric variant of the Kling-Gupta efficiency

Model calibrations with R_{KG} and R_{NP} resulted both in reasonable hydrograph simulations (Fig. 4.3) with the difference that simulations based on R_{KG} generally resulted in a wider uncertainty band than simulations based on R_{NP}. This difference was especially pronounced during low-flow periods and event recessions. However, simulations of exceptionally high flow magnitudes were better constrained for calibrations with R_{KG}. While it was the Spearman rank correlation that reduced uncertainty of low-flow simulations, it was the Pearson correlation that better constrained parameters for high-flow conditions.

Non-parametric formulations of the variability and correlation components of R_{KG} affected simulations of the various statistical metrics and signatures differently (Fig. 4.4). High-flow related hydrograph aspects (R_{eff}, R_{eff_peak}, B_{hf}) were better simulated for model calibrations with R_{KG} than R_{NP}. This negative effect of using non-parametric formulations on high flows could be predominantly attributed to the use of the Spearman rank correlation, whereas the use of the FDC did not significantly change estimates of these metrics and signatures. In contrast to high flows, estimates of low-flow related hydrograph aspects (R_{eff_logQ}, R_{MARE}, B_{lf}) were strongly improved by the use of non-parametric formulations of variability and dynamics, especially when adapting both components simultaneously (R_{NP}). The more generic signatures of runoff ratio

and watershed lag time were not much affected by the selection of the objective function, making R_{NP} and R_{KG} equivalent metrics.

Within this study it was also demonstrated that most statistical metrics and signatures were best simulated when addressing multiple hydrograph aspects (volume, variability, and dynamics) during calibration. Reducing the number of components meant losing an essential information of catchment runoff response. Especially the loss of the information on dynamics strongly impaired model performance in case of the two-component objective function. Calibration on a single component was best when focusing on dynamics, whereas calibration on volume only poorly constrained model parameters.

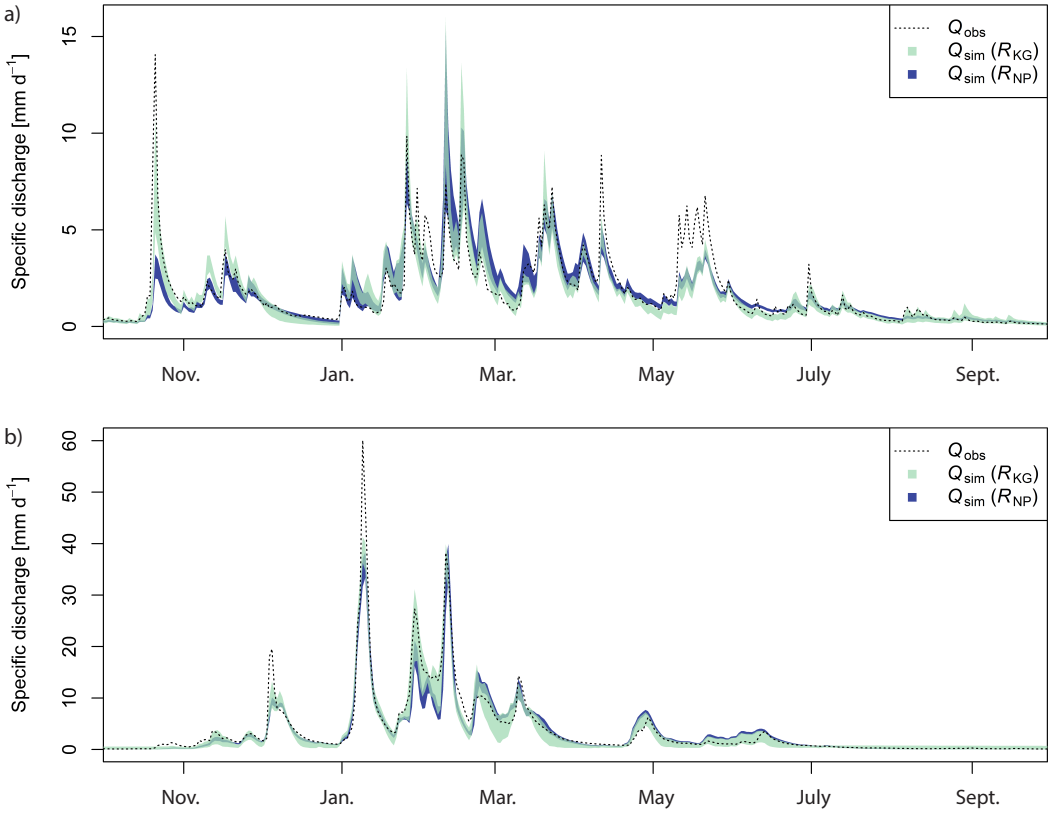


Figure 4.3: Observed (Q_{obs}) and simulated hydrographs (Q_{obs}) from model calibrations with R_{KG} and R_{NP} for a) a snow dominated catchment in the Northeast (USGS gauge id 01423000) and b) a winter-rain dominated catchment in the Northwest (USGS gauge id 14301000) of the United States. The range in hydrograph simulations indicates the 5th to 95th quantile of all 100 simulations.

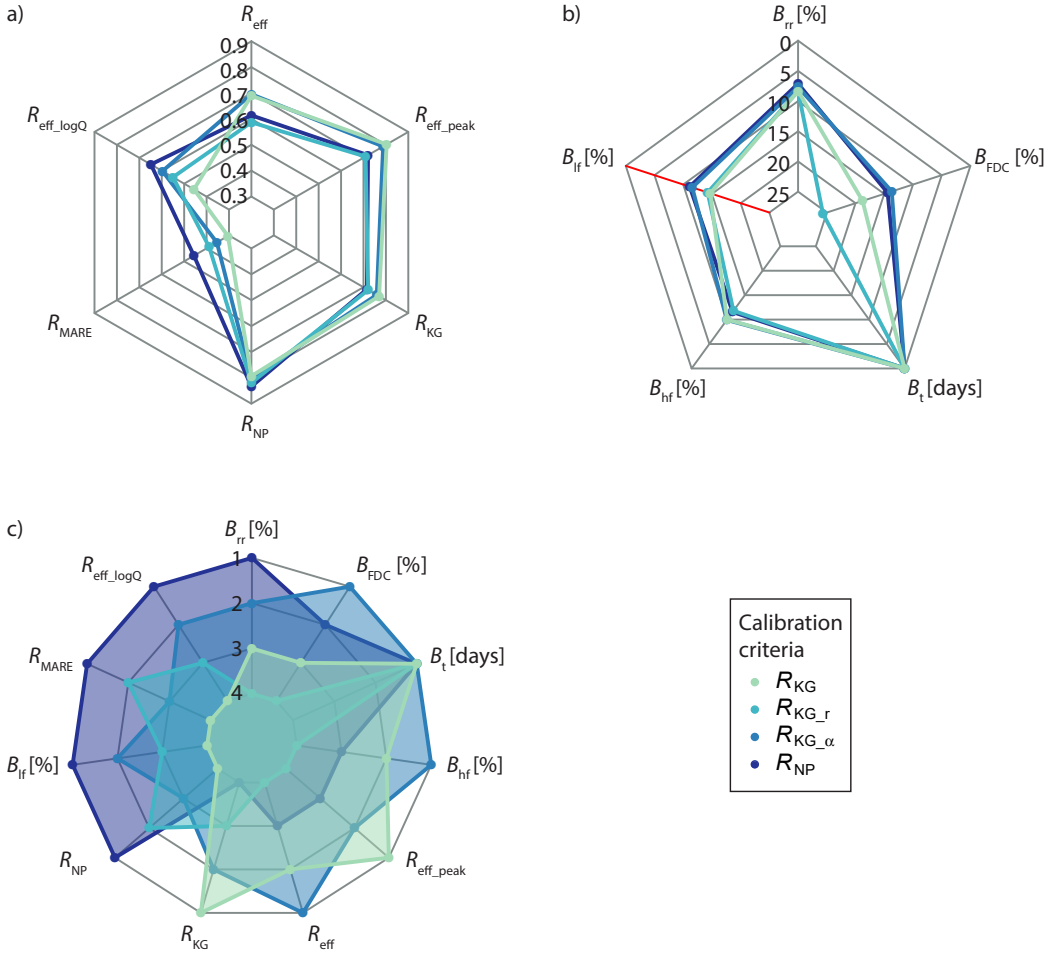


Figure 4.4: Model efficiencies in validation for model calibrations with R_{KG}, R_{KG_r}, R_{KG_α}, and R_{NP}. Calibration criteria are evaluated in terms of a) statistical metrics and b) signatures (note that the axis for B_{lf} is scaled by a factor of five meaning that percent bias is five times higher than indicated). Each calibration criteria is ranked according to its performance for statistical metrics and the signatures in c). Values correspond to the median performance of the 100 study catchments.

4.2 Gauging the ungauged catchment: Value of single discharge observations

4.2.1 Which discharge observations are most informative for model calibration?

Taking 12 discharge observations in an 'ungauged catchment' clearly improved model simulations compared to a fully uninformed situation (Fig. 4.5). Model calibration with 12 observations reached on average performance values up to 70 % and 90 % of those of a well-informed situation for the hydrograph and the FDC, respectively.

The information provided by observations from different sampling strategies had a varying value when evaluated for the hydrograph or the FDC. Hydrograph simulations were best when they were based on sampling strategies that collected information during peak flows and event recessions (e.g. $C_{Max\_Snowmelt}$). Strategies combining observations of peak flows with observations of low flows or at a fixed time interval ranked in the middle (e.g. S_{Max_Min}). Poorest model performance in terms of hydrograph efficiency was reached with discharge observations of minimum and mean discharge or discharge observations exclusively taken at fixed time intervals (e.g. S_{DOM}). The described ranking of the 13 sampling strategies was almost reversed when strategies were evaluated in terms of their information value for simulating the FDC.

The difference in the value of the 13 sampling strategies was most pronounced when HBV was calibrated with R_{eff} (Fig. 4.5a). However, changing the calibration focus by using R_{eff_logQ} as objective function (Fig. 4.5b) enabled compromise solutions with sampling strategies that could be informative for the prediction of both hydrographs and FDCs. These compromise strategies typically resulted in a collection of samples covering the full range of a catchment's discharge magnitudes by combining observations of maximum discharge with observations of minimum discharge or with observations at a fixed time interval ($C_{Max\_Rec\_DOM}$, $I_{Max\_Min\_DOM}$, S_{Max_Min}, and $C_{Max\_Min\_Wetness}$).

Sampling strategies also had an effect on constraining HBV model parameters. Parameters influencing the water balance (P_{FC}, P_{BETA}, and P_{LP} in the soil routine and P_{PERC} in the groundwater routine) could be best constrained by the information of mean and low-flow observations. In contrast, model parameters defining the timing and the shape of the hydrograph (P_{UZL}, P_{K0}, P_{MAXBAS} in the groundwater and routing routine) were more similar if observations on peak flows were used for model calibration.

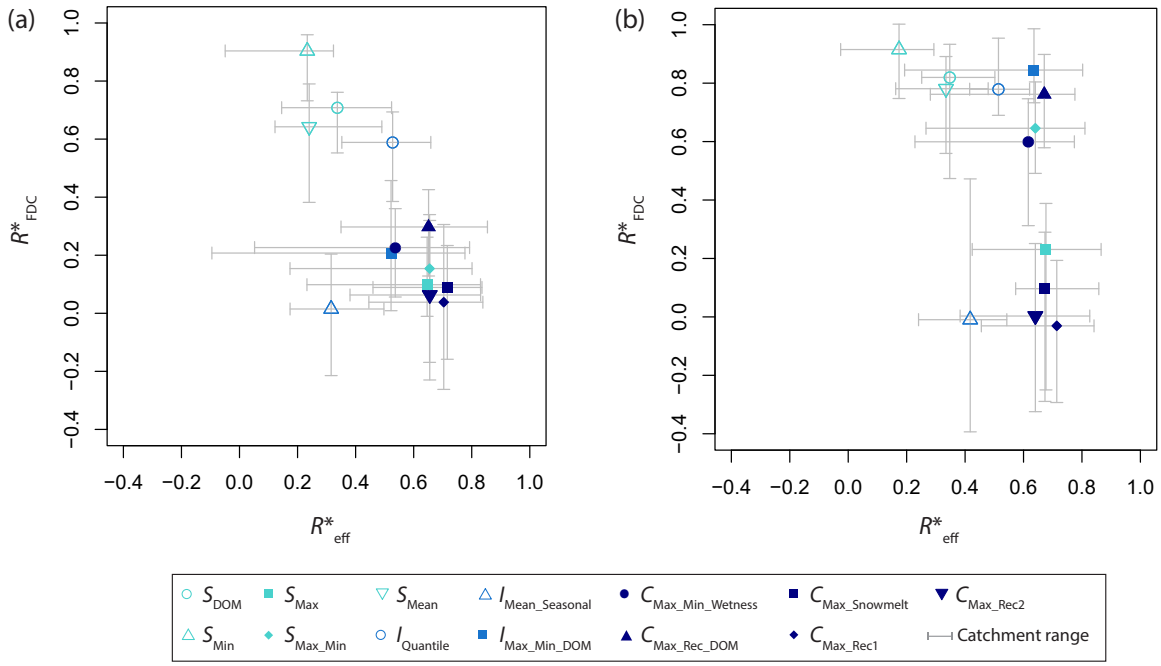


Figure 4.5: Normalized model efficiency as validated for the hydrograph (R^*_{eff}) and the FDC (R^*_{FDC}) for model calibrations with the sampling strategies using a) R_{eff} and b) R_{eff_logQ} as objective function. Each symbol represents the median model performance for a particular strategy over all catchments. It was calculated on the basis of the median ensemble mean of all calibration years. Error bars indicate the 25th to 75th quantile model performance of all catchments for the respective strategy.

4.2.2 Informing regionalization with a limited number of discharge observations

A limited number of discharge observations was generally a valuable source of information for classical regionalization. Model performance based on the information of 3 to 24 discharge observations improved classical regionalization with attribute similarity and spatial proximity by 24 % to 30 % and 22 % to 26 %, respectively. The higher effect of observations on the regionalization with attribute similarity could be assigned to the fact that spatial proximity outperformed attribute similarity in 65 % of the study catchments. Moreover, the classical spatial-proximity approach without the information of additional observations resulted in comparable efficiencies as the attribute-similarity approach informed with 24 observations.

Figure 4.6 shows that the value of 24 discharge observations varied in space. Observations were most effective in informing regionalization in arid catchments in the Southwest, in snow dominated mountainous regions or northern latitudes of the Rocky Mountains or the Atlantic Coast States, and in winter-precipitation dominated catchments typically located along the West Coast. Discharge observations had no or only limited value for regionalization in large parts of

the central region of the eastern United States, such as the Gulf Coast, the Mississippi Valley, and the Great Lakes Region.

The value of 24 discharge observations not only varied in space, but also in time, i.e., between the ten sampling years. Discharge observations collected in the most informative sampling year improved classical regionalization in 94 % (attribute similarity) and 92 % (spatial proximity) of the study catchments. However, in one or two out of ten sampling years, the 24 observations were disinformative for classical regionalization for the majority of catchments. The correlations between model efficiency and hydrometeorological conditions of a sampling year indicated that sampling years characterized by high peak discharge, or high annual or winter precipitation might be the least informative ones.

The number of discharge observations needed to effectively inform regionalization varied as a function of the sampling year. In the least informative sampling year, an increasing number of observations strongly improved regionalization with the effect that a year could change from being disinformative to being informative. In contrast, 3 discharge observations had a comparable effect on model performance as 24 discharge observations if collected in the most informative sampling year.

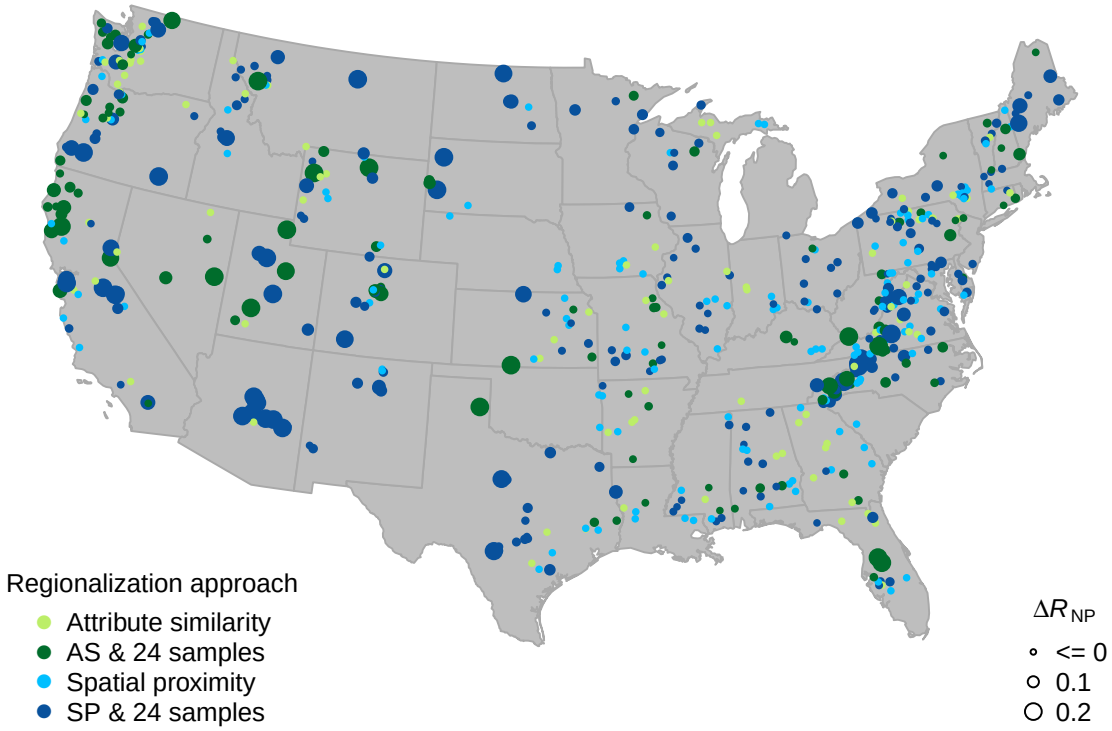


Figure 4.6: Spatial variability of the difference in validation model efficiency (ΔR_{NP}) between a classical regionalization and an informed regionalization with 24 discharge observations (efficiency of the median sampling year). The size of the circles is proportional to ΔR_{NP}, i.e., larger circles indicate a higher value of a few discharge observations for improving classical regionalization. Green circles denote catchments which were best simulated using regionalization with attribute similarity (AS), whereas blue circles indicate catchments which were best simulated using regionalization with spatial proximity (SP).

CHAPTER 5

Discussion

5.1 Value of hydrograph characteristics for model calibration

5.1.1 Prediction of ecologically relevant streamflow characteristics

SFCs are an important aspect of environmental water management (*Cartwright et al.*, 2017), because they can be related to freshwater biodiversity (for a review see *Poff and Zimmerman*, 2010). SFCs are for example used to predict the ecological state of a river after land use changes or water withdrawals (*Poff et al.*, 2010; *Olsen et al.*, 2013; *Shrestha et al.*, 2014; *Murphy et al.*, 2013). In the absence of discharge observations, SFCs have to be estimated. Accurate estimates of SFCs are essential since for high estimation errors only considerable departures from natural flow conditions can be detected (*Carlisle et al.*, 2010). Estimates of SFCs in ungauged basins have been done using statistical methods and runoff models (*Hailegeorgis and Alfredsen*, 2017). While there is a plethora of studies using multivariate regression to estimate SFCs based on catchment characteristics (e.g. *Sanborn and Bledsoe*, 2006; *Carlisle et al.*, 2010; *Knight et al.*, 2012), the use of runoff models has been limited. The regression based approach is rather inflexible in a sense that it is SFC-specific and only brings limited possibility to evaluate changes in a catchment (*Murphy et al.*, 2013). It is therefore of major interest to thoroughly evaluate and improve prediction accuracy of SFC using hydrological models (*Poff et al.*, 2010; *Olsen et al.*, 2013; *Shrestha et al.*, 2014; *Murphy et al.*, 2013).

Paper I and Paper II of this thesis contribute to the knowledge on the usefulness of hydrological models for estimating ecologically relevant SFCs. The calibration of HBV using statistical metrics or a combination of multiple statistical metrics (Paper I) resulted for the majority of SFCs in estimation accuracies of +/- 30 %, which is equivalent to the estimation uncertainty related to

the use of different observation time periods (*Kennard et al.*, 2010).

SFC estimates from calibrations with statistical metrics were usually improved when using the SFC itself as objective function (I_{Single}; Paper II). The benefit of such very accurate SFC estimates from targeted model calibration comes at the cost of a limited information value of these simulations for other SFCs. Focusing model calibration on a single SFCs probably leads to an inadequate representation of other hydrograph aspects, whereby the information value of a SFC for other SFCs can quickly become limited. For some SFCs, tailored calibrations were also related to relatively high uncertainties when moving from calibration to validation (i.e., robustness). Robustness was highest for SFCs representing physical catchment properties (e.g. recession rate or baseflow), which could be attributed to the fact that model parameters are conceptually intended to represent these characteristics. In contrast, SFCs related to high flows were the least robust, possibly because they are subject to inter-annual weather conditions and highly local precipitation dynamics. An additional indicator for the robustness of a SFCs is the length of the time series needed to calculate a SFC. For example, the two least robust SFCs were MH10 (maximum October runoff) and TL1 (timing of annual minimum runoff). Both SFCs are calculated from a single value per year. Many model parameter values can perfectly simulate this single discharge value, but a good model performance does not depend much on an accurate representation of the runoff response.

While the calibration on a single SFC improved prediction accuracy compared to a calibration on statistical metrics (i.e., R_{eff}), the use of multiple SFCs as objective function (I_{Multi}) did not make a significant difference. From these results it can be concluded that the main hydrological processes are similarly well represented with calibrations on R_{eff} and I_{Multi}. This was surprising given that I_{Multi} explicitly contained information on important hydrograph characteristics, such as baseflow, annual dischrge volume, major discharge events, and the recession rate of events. These results are in contrast to studies of *Yilmaz et al.* (2008) and *Pfannerstill et al.* (2014), where the use of multiple signatures for model calibration improved prediction accuracy for the general shape of the hydrograph. The highly variable information value of objective functions for simulating a multitude of SFCs highlights that calibration is a trade-off between finding a parameterization that is general enough to represent multiple hydrograph aspects and one that simultaneously emphasizes specific SFCs. This trade-off is a common observation as perfect model parameterizations are not possible due to a variety of uncertainty sources involved in the modeling process (*Beven*, 2016).

Concluding the discussion on the selection of objective functions for estimating SFCs, it can be argued that the variable robustness and information value of individual SFCs questions their usefulness as single objective functions, especially when using models for the prediction under changing conditions. Although I_{Multi} and R_{eff} were comparable in terms of their simulation accuracy for SFCs, I_{Multi} could be favored due to its more physically reasoned background. The decision to base multi-objective functions on SFCs requires a careful selection of signatures.

While the selection in this thesis was based on the ecological relevance of SFCs, their robustness, and information value, one could also follow the recently proposed guideline of *McMillan et al.* (2017).

From an environmental management perspective it might furthermore be of interest that results of Paper I and Paper II indicated that independent of the objective function high and mean flows were generally overestimated, whereas low flows were underestimated. This finding is in agreement with other studies evaluating simulations of ecologically relevant SFCs (*Olsen et al.*, 2013; *Kiesel et al.*, 2017; *Caldwell et al.*, 2015). The observed tendency of over and under-prediction indicates that HBV tends to retain water during event peaks and successively releases groundwater to the stream during dry periods.

5.1.2 Towards a non-parametric variant of the Kling-Gupta efficiency

The use of non-parametric objective functions is still a relatively new approach to model calibration. The comparison of simulations from calibrations with the Kling-Gupta efficiency R_{KG} and its partly non-parametric formulation R_{NP} demonstrated the value of non-parametric metrics for runoff model calibration.

Expressing flow variability in terms of the normalized FDC instead of the standard deviation positively affected simulations of all evaluated hydrological signatures and statistical metrics. The favorable effect of the FDC over the standard deviation is encouraging although it might not be surprising. It is likely related to the fact that the FDC characterizes the distribution of discharge over the full range of discharge magnitudes (*Vogel and Fennessey*, 1995), which makes it a more representative metric for flow variability than the standard deviation.

Describing discharge dynamics by the Spearman rank correlation instead of the Pearson correlation had a varied effect on discharge simulations. Since Spearman rank correlation uses ranked discharge time series it has a low sensitivity to outliers (*Krause et al.*, 2005; *Legates and McCabe*, 1999) and is linked to a loss of information. Ranked discharge time series shift the focus of model calibration away from peaks towards mean and low flows. As a consequence, mean and low-flow related metrics were well simulated with Spearman rank correlation as objective function. However, the timing and magnitude of high flows was better simulated when Pearson correlation was used for calibration. Although a reduced sensitivity of model calibrations to high flows might seem a disadvantage for certain hydrograph aspects, it makes calibration less sensitive to potential rating curve uncertainties, which are typically large for exceptional peak flow events (*McMillan et al.*, 2012). The results of Paper III also indicated that the use of Spearman rank correlation leads to a purer characterization of discharge dynamics as opposed to Pearson correlation that is a measure of both dynamics and magnitude.

The third component of R_{KG} and R_{NP}, the volume error, was described by the bias in mean discharge. There were two reasons to use the mean instead of the non-parametric median for characterizing discharge volumes. First, a hydrological model needs information on the total

discharge volume in a hydrological year to keep the water balance closed during calibration, i.e., to constrain model parameters of actual evaporation. In case of skewed distributions, the median only provides the information on the 50th quantile, whereas the mean contains information on the central tendency and the distribution. Using the median as an approximation for discharge volume would retain essential information from calibration. Second, the median discharge of semi-arid and arid catchments with prolonged dry periods can be zero, which would result in numerical problems when calculating bias metrics.

The objective function R_{NP} suggested in Paper III is a way towards calibration criteria with more realistic assumptions about the statistical nature of discharge observations and model errors. R_{NP} is therefore an interesting alternative to commonly used statistical metrics. The potential of R_{NP} is furthermore supported by the observation that model calibration with non-parametric criteria resulted in good simulations for multiple hydrograph characteristics. If timing and magnitude of high-flows are of major interest, the modeler has to be aware of the limitations of R_{NP}. However, when interested in mean and low-flows, R_{NP} could provide an alternative to the common practice of log-transformation of discharge data, which should be avoided when calculating R_{KG} (*Santos et al.*, 2018).

5.1.3 Synthesis

In Papers I to III it was demonstrated that the selection of the objective function is a critical step in the model application and should be a well reflected process. Results of Papers II and III also highlighted the importance of calibrating a runoff model against multiple hydrograph aspects that represent important aspects of a catchment's runoff response. These results are in line with the paradigm of multi-objective calibration. The use of multiple criteria should prevent an overfitting of model parameters to single hydrograph aspects and is thus expected to result in more reliable discharge predictions (some early studies are *Lindström et al.*, 1997; *Gupta et al.*, 1998; *Boyle et al.*, 2000). Evaluating model simulations against multiple criteria is also seen as a more rigorous test for runoff models (*Boyle et al.*, 2000).

While the number of criteria used for calibration probably depends on the number of model parameters (*Efstratiadis and Koutsoyiannis*, 2010), the major question to be answered is which metric to use for a multi-objective function. Should a modeler select hydrological signatures that have a very specific function, such as the ecologically relevant SFCs used in Paper II? Or should one rather combine classical statistical metrics in a way that they account for the statistical nature of discharge observations and model errors as done in Paper III? The answer to these questions is likely rather individual since it depends on a hydrologists background and modeling philosophy, but also on the ultimate aim of the modeling study.

5.2 Gauging the ungauged catchment: Value of single discharge observations

5.2.1 Which discharge observations are most informative for model calibration?

Hydrograph and FDC belong to the most widely used hydrological signatures. While the hydrograph is probably the most complex signature measuring discharge dynamics and magnitudes, the FDC only characterizes the distribution of discharge magnitudes. As shown in Paper IV, the prediction of these two signatures generally requires different flow information, i.e., observations from sampling strategies have a different value when evaluated for the FDC or the hydrograph. Results in Paper IV indicated that there are mainly two reasons for the varying value of data. One reason is the range of discharge magnitudes covered by a sampling strategy and a second reason is related to model parameters active at the time discharge samples are provided to the model. Model parameters of the groundwater and the routing routine define the shape and timing of a hydrograph. These parameters were best constrained by sampling strategies measuring peak flows and event recessions. This is likely why high-flow oriented strategies were finally the most valuable ones for the prediction of hydrographs. Results of Paper IV are therefore in agreement with previous studies reporting a relatively high value of maximum flows and recession data (*Seibert and Beven*, 2009; *Seibert and McDonnell*, 2015) or wet periods (*Yapo et al.*, 1996; *Vrugt et al.*, 2006; *Melsen et al.*, 2014) for the prediction of hydrographs compared to low flows or observations during dry periods. In contrast to hydrographs, an accurate prediction of FDCs relies on a well-modeled water balance. The water balance is to a large degree defined by parameters in the soil routine of HBV, where evaporation and groundwater recharge are calculated. To model the water balance correctly, HBV needs information on annual discharge volumes. This information is sampled by strategies that cover the full range of runoff magnitudes and result in a collection of samples with a comparable discharge distribution as continuous time series.

Although a runoff model generally needs different information for the simulation of FDCs and hydrographs, informative sampling strategies were found for both signatures if model calibration was based on the objective function R_{eff_logQ} instead of R_{eff}. Calibration with R_{eff_logQ} emphasizes mean and low flows giving more weight to a range of magnitudes and reducing the importance of accurately simulating the magnitude and timing of high flows. The shift in calibration focus had the effect that strategies could be found that were informative for both hydrograph and FDC simulations. This result confirms the findings made in Papers I to III, and highlights that a careful choice of the objective function might be even more critical when predicting discharge in data scarce regions.

5.2.2 Informing regionalization with a limited number of discharge observations

The information of a few discharge observations could considerably improve discharge predictions of classical regionalization approaches. Results of Paper V therefore confirm the findings of *Rojas-Serna et al.* (2016) and *Viviroli and Seibert* (2015), where randomly selected discharge observations or observations during mean-flow conditions were a valuable source of information for regionalization. Results of Paper V furthermore revealed that the value of observations was higher for regionalization with attribute similarity than for the spatial-proximity based regionalization approach. The difference in value was possibly due to the lower model performance of regionalization with attribute similarity than with spatial proximity. The selection of donor catchments based on common catchment attributes could lead to surprisingly large distances between donor and receiver catchment. In some cases, such as arid catchments, close catchments are likely more representative for dominant runoff processes than seemingly similar catchments from humid regions far away.

Independent of the regionalization approach, observations were most informative in arid catchments, snow dominated catchments, and catchments with winter-precipitation. These are all catchments with pronounced high-flow periods. Discharge observations used to inform regionalization were taken during these hydrologically active and important periods and therefore informed regionalization with data representing the major aspects of a runoff regime. These results are comparable to those of *Viviroli and Seibert* (2015), who reported that observations were especially valuable for snow and icemelt dominated catchments, whereas observations were less effective in constraining regionalization uncertainty in precipitation dominated catchments of Switzerland. Variability in precipitation events and the related randomness in the yearly runoff regime could be a reason for the smaller effect of observations in precipitation driven catchments (*Viviroli and Seibert*, 2015). The characteristic runoff response of these catchments possibly needs to be represented with a larger number of discharge observations.

Sampling years with high sums of annual or winter precipitation and therefore high annual peak discharge were the least informative ones for regionalization. This was unexpected at first given that observations during peak discharge (*Paper IV; Seibert and McDonnell*, 2015) and wet periods (*Yapo et al.*, 1996; *Vrugt et al.*, 2006; *Melsen et al.*, 2014) have been shown to be valuable for model calibration. However, different processes might dominate runoff response during unusually high flow events than in an average year. Observations taken during such exceptional conditions could thus be of limited representativeness for the typically observed catchment behavior. The reduced information value of observations from disinformative years could also be a reason why taking 24 observations instead of 3 was important for the least informative year, whereas 3 observations could be of comparable value as 24 when sampled in the most informative year.

5.2.3 Synthesis

Paper IV and Paper V contribute to the PUB discussion on the value of discharge data in constraining and reducing predictive uncertainty in ungauged catchments. The presented results clearly demonstrated the distinct value of a small number of observations for either directly constraining model parameters or for evaluating regionalized parameter sets.

Both studies have been based on the assumption that discharge observations are collected within a single hydrological year. While this assumption is a valid and also a realistic constraint from a practical perspective, it made the value of discharge observations sensitive to the sampling year. In Paper V, this sensitivity to sampling years could be related to the inter-annual variability in hydroclimatic conditions. The insight in what makes an informative or a disinformative year could only be gained by the use of all 579 catchments and was not discovered in Paper IV with 20 catchments. Similarly, relationships between the value of data and catchment types could only be established on the base of all 579 catchments. This highlights the potential and opportunities that large-sample data sets provide for the hydrological modeling community.

In practice, one does not know the hydrological conditions of a sampling year beforehand, i.e., is unknown if it will be an informative year. However, a crude estimate could be done at the end of the sampling year based on field observations when taking discharge measurements or by taking neighboring gauged catchments as a reference. Such an estimate would allow to indirectly assess the relative value of the observations taken during that particular year. Likewise, making measurements exactly at the annual or monthly peak discharge is a rather difficult task. The fact that there was no single best strategy but rather a set of best strategies containing similar observations is an indicator that there is some flexibility in taking discharge samples. Probably more important than the exact sampling strategy, is the active decision on the ultimate aim of discharge simulations as indicated by the findings of Papers I to IV.

Finally, in both Paper IV and Paper V modeling results were evaluated in relation to benchmarks instead of absolute performance values. Since absolute model performance can vary strongly between catchments, benchmarks are a way to evaluate model performance relative to what could be achieved at best or what should be achieved at minimum (*Seibert et al.*, 2018). The selection of benchmarks can for example depend on the modeling aim or on data availability. The concept of benchmarks is especially valuable in the context of PUB, where it is of major interest to evaluate the value of a small number of observations compared to a non-informed situation or efforts related to the maintenance of long-term gauging stations. The absolute model performance becomes more important in practical applications as soon as model performance is too low to draw reasonable conclusions from discharge simulations.

CHAPTER 6

Conclusions

In this thesis, the value of discharge data in hydrological modeling was evaluated. The major research questions were motivated by the need of accurate estimates of ecologically relevant SFCs and reliable predictions of discharge in ungauged basins. The individual studies building up this thesis therefore focused on i) the value of SFCs and (non-parametric) statistical metrics for hydrological model calibration aiming at accurate simulations of multiple hydrograph aspects, and ii) the value of a limited number of discharge observations for model calibration and informing regionalization in ungauged basins. The main findings of this thesis can be summarized as follows:

- **Value of SFCs for model calibration:** The use of SFCs as objective functions allows to base model calibration on metrics with physical meaning or ecohydrological relevance. In this thesis it was demonstrated that the variable robustness and information value of SFCs when used as objective functions requires a careful selection of SFCs. Model calibration with a single SFC or multiple SFCs is prone to the classical calibration trade-off situation, which makes accurate estimates of a diverse set of SFCs from a single hydrograph a challenging modeling task. The specification of a clear simulation aim enables targeting model calibration and in case of ungauged catchments, strategically selecting discharge observations.

- **Value of non-parametric objective functions for model calibration:** The selection of objective functions is implicitly linked to assumptions about the statistical nature of discharge data and model simulation errors. Results of this thesis demonstrated that calibration metrics accounting for non-normality, outliers, and non-linearity in data can provide an interesting alternative to commonly used statistical metrics. More specifically, the proposed modification of the popular Kling-Gupta model efficiency towards a non-parametric

metric (using the Spearman rank correlation and the flow-duration curve) resulted, except for the timing and magnitude of exceptionally high flows, in good simulations for a range of hydrograph characteristics.

- **Value of single discharge observations for model calibration:** Prediction of discharge in ungauged basins is one of the major challenges in hydrology. Results of this thesis indicated that yearly discharge sampling campaigns can improve the basis for decision making in such ungauged basins. A strategic selection of sampling days is essential given that the timing of the most informative observations depended on the ultimate simulation aim. For example, the prediction of hydrographs required information about peak flows and event recessions, whereas information on the full range of a catchment's discharge magnitudes was needed for an accurate prediction of the flow-duration curve. Within this thesis it was furthermore demonstrated that a limited number of discharge observations can be highly valuable for informing regionalization, especially in catchments with a pronounced (seasonal) runoff response.

- **Value of data and large-sample data sets:** The value of single discharge observations, the selection of SFCs or statistical metrics for model calibration, and the focus in model evaluation are inherently related to each other. Large-sample data sets support the exploration of such aspects and relationships, allow obtaining generalized results, and provide new opportunities for hydrological modeling.

CHAPTER 7

Future research

The findings of this thesis highlighted opportunities and challenges in hydrological modeling that are related to the use of single discharge observations or specific hydrograph characteristics. The practical implications of the results are twofold. First, a small number of single discharge observations can already be valuable for constraining model parameters, which encourages to take the effort of gauging ungauged catchments. Second, carefully reflecting on the modeling aim is essential since this helps to adjust model calibration to important hydrograph characteristics and to the availability of data. The findings of this thesis also raise new questions for further research about the value of different types of data, the prediction in ungauged basins, and the uncertainty of hydrological signatures.

Value of different types of data: In this thesis, the value of data for the prediction in ungauged basins was evaluated under the assumption that a hydrologist gets the opportunity to make a few discharge measurements in the otherwise ungauged catchment. In practice, a water level logger could easily be installed at the first field visit to complement discharge measurements. Continuous water level records have been shown to be surprisingly valuable for model calibration (*Seibert and Vis*, 2016). Combining information of single discharge observations and continuous water levels would allow to calibrate the model not only for volume errors, but also for discharge dynamics. Discharge uncertainty indicated by the spread in discharge simulations for a given day, could then be used to decide at which water level additional discharge measurements could be made. While in situ measurements of e.g. evaporation, soil moisture, snow, or water storage variation usually require more resources in from of equipment and labor than discharge measurements, remote sensing data of these various water balance components are becoming available (*Montanari et al.*, 2013). Such data could be used in addition to in situ observations to

evaluate model realism (i.e. internal model variables) or to evaluate the representativeness of regionalized parameter sets.

Catchment similarity and regionalization: The question about the hydrological functioning of catchments is one of the most fundamental ones in hydrology. Most regionalization approaches are based on the concept that a catchment's runoff response is tightly linked to measurable catchment characteristics. However, regionalization approaches that make explicitly use of this concept, such as parameter regression or attribute similarity, have often not performed as expected (see e.g. *Parajka et al.*, 2013). More or new knowledge on major runoff response drivers can be a crucial step in improving prediction in ungauged basins. This knowledge could be gained through comparative field experiments, but also from large-sample data sets. Large-sample data sets have only recently become available and have been used by *Berghuijs et al.* (2014) or *Kuentz et al.* (2017) to explore catchment similarity by jointly evaluating the information of catchment runoff response and attributes. Such historic data sets are seen as one of the most important sources for new information and should be more exploited (*Montanari et al.*, 2013) also with the focus on different temporal and spatial scales. In regionalization, more detailed knowledge on catchment similarity could be used to select more representative donor catchments or to evaluate regionalized parameter sets against an expected range of runoff responses. It might also be worth to explore whether runoff response and thus groups of similar catchments can be related to model structure, which would ultimately allow to regionalize model structures.

Uncertainty of hydrological signatures: Hydrological signatures are useful indices for evaluating hydrograph simulations. However, even signatures derived from observed time series can be related to considerable uncertainties due to their sensitivity to the length of a time series (*Kennard et al.*, 2010) or rating curve uncertainties (*Westerberg et al.*, 2016). Predicting signatures with hydrological models for ungauged catchments or for changing catchment conditions will add additional uncertainties to the estimated signatures. Quantifying and communicating this prediction uncertainty for a multitude of signatures and hydroclimates is essential to support a sustainable water management in the future.

Acknowledgements

Many people have contributed to this thesis by sharing their ideas and opinions and by making these four years a highly enjoyable time. I would like to specifically thank the following people:

- First I would like to thank Jan Seibert, my main supervisor, for giving me the opportunity to do a PhD in his group, joining many summer schools, presenting at conferences, and even visiting a small sample of my 579 study catchments in the United States. I learnt a lot from Jan and always enjoyed the critical, fruitful, respectful, open minded, and inspiring discussions we had.
- Thanks to Daniel Viviroli, my second supervisor, for sharing his time, ideas, and flair for details with me.
- Thanks to Christian Huggel for the input during the committee meetings.
- I thank Marc Vis and Tracy Ewen, my two office colleagues during the first half year of my PhD, for answering the many questions on HBV, the US catchment dataset, and statistics, as well as for the helpful and prompt language-checks of the manuscripts before submission.
- Thanks to my colleagues from H2K for the enjoyable company during the many coffee breaks, lunch breaks, retreats and H2K-excursions.
- This PhD project was funded by the University of Zurich (Canton of Zurich), which is gratefully acknowledged. The Graduate School of the Geography Department supported me with travel funding.
- Thanks to my two friends Anna Sikorska-Senoner and Ling Wang for the joyful atmosphere in our office, all the scientific but more importantly personal chats, watering our jungle plants, feedback on posters and presentations, sharing the apartment during conferences, birthday cards, etc.
- And last, many thanks to my family and Mark for supporting me in anything I would like to do and for setting the foundation for my curiosity about the world.

References

Abell, R., D. Olson, E. Dinerstein, P. Hurley, J. Diggs, W. Eichbaum, S. Walters, W. Wettengel, T. Allnutt, C. Loucks, and P. Hedao (2000), *Freshwater Ecoregions of North America: A Conservation Assessment*, Island Press, Washington, DC, USA.

Addor, N., A. J. Newman, N. Mizukami, and M. P. Clark (2017), The CAMELS data set: Catchment attributes and meteorology for large-sample studies, *Hydrology and Earth System Sciences*, *21*(10), 5293–5313, doi:10.5194/hess-21-5293-2017.

Arsenault, R., and F. Brissette (2014), Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches, *Water Resources Research*, (50), 6135–6153, doi:10.1002/2013WR014898.

Arthington, A. H., S. E. Bunn, N. L. Poff, and R. J. Naiman (2006), The challenge of providing environmental flow rules to sustain river ecosystems., *Ecological Applications*, *16*(4), 1311–1318, doi:10.1890/1051-0761(2006)016[1311:TCOPEF]2.0.CO;2,.

Bárdossy, A. (2007), Calibration of hydrological model parameters for ungauged catchments, *Hydrology and Earth System Sciences*, *11*(2), 703–710, doi:10.5194/hess-11-703-2007.

Berghuijs, W. R., M. Sivapalan, R. A. Woods, and H. H. H. Savenije (2014), Patterns of similarity of seasonal water balances: a window into streamflow variability over a range of time scales, *Water Resources Research*, *50*, 5638–5661, doi:10.1002/2014WR015692.

Bergström, S. (1976), Development and application of a conceptual runoff model for Scandinavian catchments, *Tech. rep.*, SMHI, Report No. RHO 7, Norrköping, Sweden.

Bergström, S. (1992), The HBV Model: Its Structure and Applications, *Tech. rep.*, SMHI, Report No. RH 4, Norrköping, Sweden.

Bergström, S., and G. Lindström (2015), Interpretation of runoff processes in hydrological modelling-experience from the HBV approach, *Hydrological Processes*, *29*(16), 3535–3545, doi:10.1002/hyp.10510.

Beven, K. (2012), *Rainfall-runoff modelling: The primer*, 2nd ed., Wiley-Blackwell, Chichester, UK.

Beven, K. (2016), Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication, *Hydrological Sciences Journal*, *61*(9), 1652–1665, doi:10.1080/02626667.2015.1031761.

Blöschl, G., and M. Sivapalan (1995), Scale issues in hydrological modelling: A review, *Hydrological Processes*, *9*, 251–290, doi:10.5194/hess-19-4559-2015.

Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods, *Water Resources Research*, *36*(12), 3663–3674, doi:10.1029/2000WR900207.

Brath, A., A. Montanari, and E. Toth (2004), Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrological model, *Journal of Hydrology*, *291*(3-4), 232–253, doi:10.1016/j.jhydrol.2003.12.044.

Burn, D. H. (1990), Evaluation of regional flood frequency analysis with a region of influence approach, *Water Resources Research*, *26*(10), 2257–2266, doi:10.1029/90WR01192.

Burn, D. H., and D. B. Boorman (1992), Catchment classification applied to the estimation of hydrological parameters at ungauged catchments, *Tech. rep.*, Instistute of Hydrology, Wallingford, UK.

Buytaert, W., and K. Beven (2009), Regionalization as a learning process, *Water Resources Research*, *45*(11), W11,419, doi:10.1029/2008WR007359.

Caldwell, P. V., J. G. Kennen, G. Sun, J. E. Kiang, J. B. Butcher, M. C. Eddy, L. E. Hay, J. H. LaFontaine, E. F. Hain, S. A. C. Nelson, and S. G. McNulty (2015), A comparison of hydrologic models for ecological flows and water availability, *Ecohydrology*, *8*, 1525–1546, doi:10.1002/eco.1602.

Carlisle, D. M., J. Falcone, D. M. Wolock, M. R. Meador, and R. H. Norris (2010), Predicting the natural flow regime: Models for assessing hydrological alteration in streams, *River Research and Applications*, *26*, 118–136, doi:10.1002/rra.1247.

Cartwright, J., C. Caldwell, S. Nebiker, and R. Knight (2017), Putting flow-ecology relationships into practice: A decision-support system to assess fish community response towatermanagement scenarios, *Water*, *9*(3), 13–16, doi:10.3390/w9030196.

Efstratiadis, A., and D. Koutsoyiannis (2010), One decade of multi-objective calibration approaches in hydrological modelling: A review, *Hydrological Sciences Journal*, *55*(1), 58–78, doi:10.1080/02626660903526292.

GRDC (2018), Global runoff data center: A repository for the world's river discharge data and associated metadata, https://www.bafg.de/GRDC/EN/Home/homepage_node.html.

Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and non commensurable measures of information, *Water Resource*, *34*(4), 751–763, doi:doi:10.1029/97WR03495.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez (2009), Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, *Journal of Hydrology*, *377*(1-2), 80–91, doi:10.1016/j.jhydrol.2009.08.003.

Hailegeorgis, T. T., and K. Alfredsen (2017), Regional statistical and precipitation–runoff modelling for ecological applications: prediction of hourly streamflow in regulated rivers and ungauged basins, *River Research and Applications*, *33*, 233–248, doi:10.1002/rra.3006.

Harlin, J. (1991), Development of a process oriented calibration scheme for the HBV hydrological model, *Nordic Hydrology*, *22*, 15–36, doi:10.2166/nh.1991.002.

He, Y., A. Bárdossy, and E. Zehe (2011), A review of regionalisation for continuous streamflow simulation, *Hydrology and Earth System Sciences*, *15*(11), 3539–3553, doi:10.5194/hess-15-3539-2011.

Hingray, B., B. Schaefli, A. Mezghani, and Y. Hamdi (2010), Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments, *Hydrological Sciences Journal*, *55*(6), 1002–1016, doi:10.1080/02626667.2010.505572.

Hrachowitz, M., H. Savenije, G. Blöschl, J. McDonnell, M. Sivapalan, J. Pomeroy, B. Arheimer, T. Blume, M. Clark, U. Ehret, F. Fenicia, J. Freer, a. Gelfan, H. Gupta, D. Hughes, R. Hut, a. Montanari, S. Pande, D. Tetzlaff, P. Troch, S. Uhlenbrook, T. Wagener, H. Winsemius, R. Woods, E. Zehe, and C. Cudennec (2013), A decade of Predictions in Ungauged Basins (PUB)–a review, *Hydrological Sciences Journal*, *58*(6), 1198–1255, doi:10.1080/02626667.2013.803183.

Hughes, D. A., M. Gush, J. Tanner, and P. Dye (2014), Using targeted short-term field investigations to calibrate and evaluate the structure of a hydrological model, *Hydrological Processes*, *28*(5), 2794–2809, doi:10.1002/hyp.9807.

Jarvis, A., H. Reuter, A. Nelson, and E. Guevara (2008), Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m, http://srtm.csi.cgiar.org.

Juston, J., J. Seibert, and P.-O. Johansson (2009), Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment, *Hydrological Processes*, *23*(21), 3093–3109, doi:10.1002/hyp.7421.

Kennard, M. J., S. J. Mackay, B. J. Pusey, J. D. Olden, and N. Marsh (2010), Quantifying uncertainty in estimation of hydrologic metrics for ecohydrological studies, *River Research and Applications*, *26*, 137–156, doi:10.1002/rra.1249.

Kiesel, J., B. Guse, M. Pfannerstill, K. Kakouei, S. C. Jähnig, and N. Fohrer (2017), Improving hydrological model optimization for riverine species, *Ecological Indicators*, *80*, 376–385, doi: 10.1016/j.ecolind.2017.04.032.

Kim, U., and J. J. Kaluarachchi (2009), Hydrologic model calibration using discontinuous data: An example from the upper Blue Nile River Basin of Ethiopia, *Hydrological Processes*, *23*(26), 3705–3717, doi:10.1002/hyp.7465.

Klemes, V. (1986), Operational testing of hydrological simulation models, *Hydrological Sciences Journal*, *31*(1), 13–24, doi:10.1080/02626668609491024.

Knight, R. R., M. Brian Gregory, and A. K. Wales (2008), Relating streamflow characteristics to specialized insectivores in the Tennessee River Valley: A regional approach, *Ecohydrology*, *1*(4), 394–407, doi:10.1002/eco.32.

Knight, R. R., W. S. Gain, and W. J. Wolfe (2012), Modelling ecological flow regime: An example from the Tennessee and Cumberland River basins, *Ecohydrology*, *5*(5), 613–627, doi:10.1002/eco.246.

Knight, R. R., J. C. Murphy, W. J. Wolfe, C. F. Saylor, and A. K. Wales (2014), Ecological limit functions relating fish community response to hydrologic departures of the ecological flow regime in the Tennessee River basin, United States, *Ecohydrology*, *7*(5), 1262–1280, doi: 10.1002/eco.1460.

Kokkonen, T. S., A. J. Jakeman, P. C. Young, and H. J. Koivusalo (2003), Predicting daily flows in ungauged catchments: Model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina, *Hydrological Processes*, *17*(11), 2219–2238, doi:10.1002/hyp.1329.

Konz, M., and J. Seibert (2010), On the value of glacier mass balances for hydrological model calibration, *Journal of Hydrology*, *385*(1-4), 238–246, doi:10.1016/j.jhydrol.2010.02.025.

Krause, P., D. P. Boyle, and F. Bäse (2005), Comparison of different efficiency criteria for hydrological model assessment, *Advances in Geosciences*, *5*, 89–97, doi:10.5194/adgeo-5-89-2005.

Kuentz, A., B. Arheimer, Y. Hundecha, and T. Wagener (2017), Understanding hydrologic variability across Europe through catchment classification, *Hydrology and Earth System Sciences*, *21*(6), 2863–2879, doi:10.5194/hess-21-2863-2017.

Lebecherel, L., V. Andréassian, and C. Perrin (2016), On evaluating the robustness of spatial-proximity-based regionalization methods, *Journal of Hydrology*, *539*, 196–203, doi:10.1016/j.jhydrol.2016.05.031.

Legates, D. R., and G. J. McCabe (1999), Evaluating the use of 'goodness-of-fit' measures in hydrologic and hydroclimatic model validation, *Water Resources Research*, *35*(1), 233–241, doi:10.1029/1998WR900018.

Lehner, B., and P. Döll (2004), Development and validation of a global database of lakes, reservoirs and wetlands, *Journal of Hydrology*, *296*(1-4), 1–22, doi:10.1016/j.jhydrol.2004.03.028.

Lindström, G., B. Johansson, M. Persson, M. Gardelin, and S. Bergström (1997), Development and test of the distributed HBV-96 hydrological model, *Journal of Hydrology*, *201*(1-4), 272–288, doi:10.1016/S0022-1694(97)00041-3.

McIntyre, N., H. Lee, H. Wheater, A. Young, and T. Wagener (2005), Ensemble predictions of runoff in ungauged catchments, *Water Resources Research*, *41*(12), 1–14, doi:10.1029/2005WR004289.

McMillan, H., T. Krueger, and J. Freer (2012), Benchmarking observational uncertainties for hydrology: Rainfall, river discharge and water quality, *Hydrological Processes*, *26*(26), 4078–4111, doi:10.1002/hyp.9384.

McMillan, H., I. Westerberg, and F. Branger (2017), Five guidelines for selecting hydrological signatures, *Hydrological Processes*, *31*(26), 4757–4761, doi:10.1002/hyp.11300.

Melsen, L. A. L., A. J. Teuling, S. W. S. van Berkum, P. J. J. F. Torfs, and R. Uijlenhoet (2014), Catchments as simple dynamical systems: A case study on methods and data requirements for parameter identification, *Water Resources Research*, *50*, 5577–5596, doi:10.1002/2013WR014720.

Merz, R., and G. Blöschl (2004), Regionalisation of catchment model parameters, *Journal of Hydrology*, *287*(1-4), 95–123, doi:10.1016/j.jhydrol.2003.09.028.

Merz, R., J. Parajka, and G. Blöschl (2009), Scale effects in conceptual hydrological modeling, *Water Resources Research*, *45*(9), 1–15, doi:10.1029/2009WR007872.

Milligan, G. W., and M. C. Cooper (1988), A study of standardizaction of variables in cluster analysis, *Journal of Classification*, *5*, 181–204, doi:10.1007/BF01897163.

Montanari, A., G. Young, H. H. Savenije, D. Hughes, T. Wagener, L. L. Ren, D. Koutsoyiannis, C. Cudennec, E. Toth, S. Grimaldi, G. Blöschl, M. Sivapalan, K. Beven, H. Gupta, M. Hipsey, B. Schaefli, B. Arheimer, E. Boegh, S. J. Schymanski, G. Di Baldassarre, B. Yu, P. Hubert, Y. Huang, A. Schumann, D. A. Post, V. Srinivasan, C. Harman, S. Thompson, M. Rogger, A. Viglione, H. McMillan, G. Characklis, Z. Pang, and V. Belyaev (2013), "Panta Rhei-Everything Flows": change in hydrology and society-the IAHS Scientific Decade 2013-2022, *Hydrological Sciences Journal*, *58*(6), 1256–1275, doi:10.1080/02626667.2013.809088.

Murphy, A. (1988), Skill scores based on the mean square error and their relationships to the correlation coefficient, *Monthly Weather Review*, *116*, 2417–2424.

Murphy, J. C., R. R. Knight, W. J. Wolfe, and W. S. Gain (2013), Predicting ecological flow regime at ungaged sites: A comparison of methods, *River Research and Applications*, *29*(5), 660–669, doi:10.1002/rra.2570.

Nash, J. E., and J. V. Sutcliffe (1970), River flow forecasting through conceptual models part I - A discussion of principles, *Journal of Hydrology*, *10*(3), 282–290, doi:10.1016/0022-1694(70)90255-6.

Newman, A. J., M. P. Clark, K. Sampson, A. Wood, L. E. Hay, A. Bock, R. J. Viger, D. Blodgett, L. Brekke, J. R. Arnold, T. Hopson, and Q. Duan (2015), Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance, *Hydrology and Earth System Sciences*, *19*(1), 209–223, doi:10.5194/hess-19-209-2015.

NOAA (2018), Geographical reference maps, https://www.ncdc.noaa.gov/temp-and-precip/drought/nadm/geography.

Olden, J. D., and N. L. Poff (2003), Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, *River Research and Applications*, *19*(2), 101–121, doi:10.1002/rra.700.

Olsen, M., L. Troldborg, H. J. Henriksen, J. Conallin, J. C. Refsgaard, and E. Boegh (2013), Evaluation of a typical hydrological model in relation to environmental flows, *Journal of Hydrology*, *507*, 52–62, doi:10.1016/j.jhydrol.2013.10.022.

Oudin, L., V. Andréassian, C. Perrin, C. Michel, and N. Le Moine (2008), Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments, *Water Resources Research*, *44*(3), 1–15, doi:10.1029/2007WR006240.

Parajka, J., A. Viglione, M. Rogger, J. L. Salinas, M. Sivapalan, and G. Blöschl (2013), Comparative assessment of predictions in ungauged basins – Part 1: Runoff-hydrograph studies, *Hydrology and Earth System Sciences*, *17*(5), 1783–1795, doi:10.5194/hess-17-1783-2013.

Perrin, C., L. Oudin, V. Andreassian, C. Rojas-Serna, C. Michel, and T. Mathevet (2007), Impact of limited streamflow data on the efficiency and the parameters of rainfall–runoff models, *Hydrological Sciences Journal*, *52*(1), 131–151, doi:10.1623/hysj.52.1.131.

Pfannerstill, M., B. Guse, and N. Fohrer (2014), Smart low flow signature metrics for an improved overall performance evaluation of hydrological models, *Journal of Hydrology*, *510*, 447–458, doi:10.1016/j.jhydrol.2013.12.044.

Pfannerstill, M., K. Bieger, B. Guse, D. D. Bosch, N. Fohrer, and J. G. Arnold (2017), How to constrain multi-objective calibrations of the SWAT model using water balance components,

Journal of the American Water Resources Association, *53*(3), 532–546, doi:10.1111/1752-1688.12524.

Poff, N., B. Richter, A. Arthington, S. Bunn, R. Naiman, E. Kendy, M. Acreman, C. Apse, B. Bledsoe, M. Freeman, J. Henriksen, R. Jacobsen, J. Kennen, D. Merritt, J. O'Keeffe, J. Olden, K. Rogers, R. Tharme, Warne, and A. (2010), The ecological limits of hydrologic alteration (ELOHA): A new framework for developing regional environmental flow standards, *Freshwater Biology*, *55*(1), 147–170, doi:10.1111/j.1365-2427.2009.02204.x.

Poff, N. L., and J. K. Zimmerman (2010), Ecological responses to altered flow regimes: A literature review to inform the science and management of environmental flows, *Freshwater Biology*, *55*(1), 194–205, doi:10.1111/j.1365-2427.2009.02272.x.

Poff, N. L., J. D. Allan, M. B. Bain, J. R. Karr, K. L. Prestegaard, B. D. Richter, R. E. Sparks, and J. C. Stromberg (1997), A paradigm for river conservation and restoration, *BioScience*, *47*(11), 769–784, doi:10.2307/1313099.

Priestley, C. H. B., and R. J. Taylor (1972), On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters, *Monthly Weather Review*, *100*(2), 81–92, doi:10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2.

Razavi, T., P. Coulibaly, and M. Asce (2013), Streamflow prediction in ungauged basins: Review of regionalization methods, *Journal of Hydrologic Engineering*, *18*(8), 958–975, doi:10.1061/(ASCE)HE.1943-5584.0000690.

Richter, B. D., J. V. Baumgartner, J. Powell, and D. P. Braun (1996), A method for assessing hydrologic alteration within ecosystems, *Conservation Biology*, *10*(4), 1163–1174, doi:10.2307/2387152.

Rode, M., U. Suhr, and G. Wriedt (2007), Multi-objective calibration of a river water quality model– Information content of calibration data, *Ecological Modelling*, *204*(1-2), 129–142, doi:10.1016/j.ecolmodel.2006.12.037.

Rojas-Serna, C., L. Lebecherel, C. Perrin, V. Andréassian, and L. Oudin (2016), How should a rainfall-runoff model be parameterized in an almost ungauged catchment? A methodology tested on 609 catchments, *Water Resources Research*, *52*, 4765–4784, doi:10.1002/2015WR018549.

Rotstayn, L. D., M. L. Roderick, and G. D. Farquhar (2006), A simple pan-evaporation model for analysis of climate simulations: Evaluation over Australia, *Geophysical Research Letters*, *33*(17), L17,715, doi:10.1029/2006GL027114.

Ryo, M., Y. Iwasaki, C. Yoshimura, and O. C. Saavedra V. (2015), Evaluation of spatial pattern of altered flow regimes on a river network using a distributed hydrological model, *PLoS ONE*, *10*(7), 1–16, doi:10.1371/journal.pone.0133833.

Samuel, J., P. Coulibaly, and R. A. Metcalfe (2011), Estimation of continuous streamflow in Ontario ungauged basins: Comparison of regionalization methods, *Journal of Hydrologic Engineering*, *16*(5), 447–459, doi:10.1061/(ASCE)HE.1943-5584.0000338.

Sanborn, S. C., and B. P. Bledsoe (2006), Predicting streamflow regime metrics for ungauged streamsin Colorado, Washington, and Oregon, *Journal of Hydrology*, *325*(1-4), 241–261, doi:10.1016/j.jhydrol.2005.10.018.

Santos, L., G. Thirel, and C. Perrin (2018), Technical note: Pitfalls in using log-transformed flows within the KGE criterion, *Hydrology and Earth System Sciences Discussions*, pp. 1–14, doi:10.5194/hess-2018-298.

Schaefli, B., and H. V. Gupta (2007), Do Nash values have value?, *Hydrological Processes*, *21*(15), 2075–2080, doi:10.1002/hyp.6825.

Seibert, J. (1999), Regionalisation of parameters for a conceptual rainfall-runoff model, *Agricultural and Forest Meteorology*, *98-99*, 279–293, doi:10.1016/S0168-1923(99)00105-7.

Seibert, J. (2000), Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, *Hydrology and Earth System Sciences*, *4*(2), 215–224, doi:10.5194/hess-4-215-2000.

Seibert, J., and K. J. Beven (2009), Gauging the ungauged basin: How many discharge measurements are needed?, *Hydrology and Earth System Sciences*, *13*(6), 883–892, doi:10.5194/hess-13-883-2009.

Seibert, J., and J. J. McDonnell (2002), On the dialog between experimentalist and modeler in catchment hydrology: Use of soft data for multicriteria model calibration, *Water Resources Research*, *38*(11), 23–1–23–14, doi:10.1029/2001WR000978.

Seibert, J., and J. J. McDonnell (2015), Gauging the ungauged basin: Relative value of soft and hard data, *Journal of Hydrologic Engineering*, *20*(1), A4014,004, doi:10.1061/(ASCE)HE.1943-5584.0000861.

Seibert, J., and M. Vis (2016), How informative are stream level observations in different geographic regions?, *Hydrological Processes*, *30*(14), 2498–2508, doi:10.1002/hyp.10887.

Seibert, J., and M. J. P. Vis (2012), Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, *Hydrology and Earth System Sciences*, *16*(9), 3315–3325.

Seibert, J., M. Vis, E. Lewis, and H. J. van Meerveld (2018), Upper and lower benchmarks in hydrological modelling, *Hydrological Processes*, *32*(8), 1120–1125, doi:10.1002/hyp.11476.

Shafii, M., N. Basu, J. R. Craig, S. L. Schiff, and P. Van Cappellen (2017), A diagnostic approach to constraining flow partitioning in hydrologic models using a multiobjective optimization framework, *Water Resources Research*, *53*(4), 3279–3301, doi:10.1002/2016WR019736.

Shrestha, R. R., D. L. Peters, and M. a. Schnorbus (2014), Evaluating the ability of a hydrologic model to replicate hydro-ecologically relevant indicators, *Hydrological Processes*, *28*(14), 4294–4310, doi:10.1002/hyp.9997.

Singh, S. K., and A. Bárdossy (2012), Calibration of hydrological models on hydrologically unusual events, *Advances in Water Resources*, *38*, 81–91, doi:10.1016/j.advwatres.2011.12.006.

Sivapalan, M., K. Takeuchi, S. W. Franks, V. K. Gupta, H. Karambiri, V. Lakshmi, X. Liang, J. J. McDonnell, E. M. Mendiondo, P. E. O'Connell, T. Oki, J. W. Pomeroy, D. Schertzer, S. Uhlenbrook, E. Zehe, P. O'Connell, T. Oki, J. W. Pomeroy, D. Schertzer, S. Uhlenbrook, E. Zehe, P. E. O'Connell, T. Oki, J. W. Pomeroy, D. Schertzer, S. Uhlenbrook, and E. Zehe (2003), IAHS Decade on Predictions in Ungauged Basins (PUB), 2003-2012: Shaping an exciting future for the hydrological sciences, *Hydrological Sciences Journal*, *48*(6), 857–880, doi:10.1623/hysj.48.6.857.51421.

Sun, W., Y. Wang, G. Wang, X. Cui, J. Yu, D. Zuo, and Z. Xu (2017), Physically based distributed hydrological model calibration based on a short period of streamflow data: Case studies in four Chinese basins, *Hydrology and Earth System Sciences*, *21*(1), 251–265, doi:10.5194/hess-21-251-2017.

Tada, T., and K. J. Beven (2012), Hydrological model calibration using a short period of observations, *Hydrological Processes*, *26*(6), 883–892, doi:10.1002/hyp.8302.

Thornton, P. E., M. M. Thornton, B. W. Mayer, N. Wilhelmi, Y. Wei, R. Devarakonda, and R. B. Cook (2014), Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2., http://www.osti.gov/scitech/biblio/1148868.

Tobler, W. R. (1970), A computer movie simulating urban growth in the Detroit Region, *Economic Geography*, *46*, 234–240, doi:10.1126/science.11.277.620.

Uhlenbrook, S., J. Seibert, C. Leibundgut, and A. Rodhe (1999), Prediction uncertainty of conceptual rainfall-runoff models caused by problems in identifying model parameters and structure, *Hydrological Sciences Journal*, *44*(5), 779–797, doi:10.1080/02626669909492273.

U.S. Department of Commerce (2007), Climatography of the United States No. 85 Divisional Normals and Standard Deviations of Temperature, Precipitation, and Heating and Cooling Degree Days 1971–2000 (And previous normals periods).

U.S. Geological Survey (2014a), USGS surface-water data for the nation, https://waterdata.usgs.gov/usa/nwis/sw.

U.S. Geological Survey (2014b), EflowStats R-package, https://github.com/USGS-R/EflowStats.

Viglione, A., J. Parajka, M. Rogger, J. L. Salinas, G. Laaha, M. Sivapalan, and G. Blöschl (2013), Comparative assessment of predictions in ungauged basins - Part 3: Runoff signatures in Austria, *Hydrology and Earth System Sciences*, *17*(6), 2263–2279, doi:10.5194/hess-17-2263-2013.

Viviroli, D., and J. Seibert (2015), Can a regionalized model parameterisation be improved with a limited number of runoff measurements?, *Journal of Hydrology*, *529*, 49–61, doi:10.1016/j.jhydrol.2015.07.009.

Vogel, R. M., and N. M. Fennessey (1995), Flow duration curves ll: A review of applications in water resources planning, *Water Resources Bulletin*, *31*, 1029–1039, doi:10.1111/j.1752-1688.1995.tb03419.x.

Vrugt, J. A., H. V. Gupta, S. C. Dekker, S. Sorooshian, T. Wagener, and W. Bouten (2006), Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting model, *Journal of Hydrology*, *325*(1-4), 288–307, doi:10.1016/j.jhydrol.2005.10.041.

Westerberg, I. K., J. L. Guerrero, P. M. Younger, K. J. Beven, J. Seibert, S. Halldin, J. E. Freer, and C. Y. Xu (2011), Calibration of hydrological models using flow-duration curves, *Hydrology and Earth System Sciences*, *15*(7), 2205–2227, doi:10.5194/hess-15-2205-2011.

Westerberg, I. K., T. Wagener, G. Coxon, H. K. McMillan, A. Castellarin, A. Montanari, and J. Freer (2016), Uncertainty in hydrological signatures for gauged and ungagauged catchments, *Water Resource Research*, *52*, 1847–1865, doi:10.1002/2015WR017635.

Xia, Y. (2004), Impacts of data length on optimal parameter and uncertainty estimation of a land surface model, *Journal of Geophysical Research*, *109*(D7), D07,101, doi:10.1029/2003JD004419.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1996), Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data, *Journal of Hydrology*, *181*(1-4), 23–48, doi:10.1016/0022-1694(95)02918-4.

Yilmaz, K. K., H. V. Gupta, and T. Wagener (2008), A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, *Water Resources Research*, *44*, W09,417, doi:10.1029/2007WR006716.

Zhang, Y., and F. H. Chiew (2009), Relative merits of different methods for runoff predictions in ungauged catchments, *Water Resources Research*, *45*, W07,412, doi:10.1029/2008WR007504.

Paper I

Water **2015**, *7*, 2358-2381; doi:10.3390/w7052358

OPEN ACCESS

water

ISSN 2073-4441

www.mdpi.com/journal/water

Article

Model Calibration Criteria for Estimating Ecological Flow Characteristics

Marc Vis [1,†,*], Rodney Knight [2,†], Sandra Pool [1], William Wolfe [2] and Jan Seibert [1,3,†]

[1] Department of Geography, University of Zurich, Winterthurerstrasse 190, 8057 Zurich, Switzerland; E-Mails: sandra.pool@geo.uzh.ch (S.P.); jan.seibert@geo.uzh.ch (J.S.)

[2] U.S. Geological Survey Lower Mississippi—Gulf Water Science Center, 640 Grassmere Park, Suite 100, Nashville, TN 37211, USA; E-Mails: rrknight@usgs.gov (R.K.); wjwolfe@usgs.gov (W.W.)

[3] Department of Earth Sciences, Uppsala University, Villavägen 16, 752 36 Uppsala, Sweden

† These authors contributed equally to this work.

* Author to whom correspondence should be addressed; E-Mail: marc.vis@geo.uzh.ch; Tel.: +41-44-635-5174.

Academic Editor: Lutz Breuer

Received: 31 January 2015 / Accepted: 4 May 2015 / Published: 20 May 2015

Abstract: Quantification of streamflow characteristics in ungauged catchments remains a challenge. Hydrological modeling is often used to derive flow time series and to calculate streamflow characteristics for subsequent applications that may differ from those envisioned by the modelers. While the estimation of model parameters for ungauged catchments is a challenging research task in itself, it is important to evaluate whether simulated time series preserve critical aspects of the streamflow hydrograph. To address this question, seven calibration objective functions were evaluated for their ability to preserve ecologically relevant streamflow characteristics of the average annual hydrograph using a runoff model, HBV-light, at 27 catchments in the southeastern United States. Calibration trials were repeated 100 times to reduce parameter uncertainty effects on the results, and 12 ecological flow characteristics were computed for comparison. Our results showed that the most suitable calibration strategy varied according to streamflow characteristic. Combined objective functions generally gave the best results, though a clear underprediction bias was observed. The occurrence of low prediction errors for certain combinations of objective function and flow characteristic suggests that (1) incorporating multiple ecological flow characteristics into a single objective

function would increase model accuracy, potentially benefitting decision-making processes; and (2) there may be a need to have different objective functions available to address specific applications of the predicted time series.

Keywords: hydrological modeling; ecological flow characteristics; objective functions; model calibration; parameter uncertainty; catchments

1. Introduction

The interactions between streamflow and aquatic ecosystems have occupied researchers across a range of disciplines for more than 50 years. Beginning with studies as early as Rantz [1] and continuing through Tennant [2] to the present day, numerous individual streamflow characteristics have been associated with various ecological responses [3]. More recently, studies have emphasized the importance of multiple streamflow characteristics operating simultaneously or interacting to influence ecological outcomes [4]. These streamflow characteristics are used to quantify relations between flow and ecological responses. At sites where streamflow records are available, the ecologically relevant streamflow characteristics (SFCs) can be derived directly from streamflow observations. However, many, probably most, sites of biological interest have few if any observed streamflow records.

Where streamflow records are unavailable, hydrological modeling is commonly used to derive flow time series, and these simulated time series are then used to derive streamflow characteristics. The basic assumption is that if a model is capable of reproducing observed streamflow with some accuracy, the simulated time series are also suitable to derive ecologically relevant flow characteristics. However, one has to note that flow simulations are never perfect and that they generally depend on the model and its parameterization. Therefore, the suitability of simulated flow series as a basis for the estimation of streamflow characteristics might vary considerably. Key issues that must be addressed include which aspects of the stream hydrograph (SFCs) should be estimated and which modeling approaches are best suited for estimating them.

At least two broad approaches to hydrologic modeling have been applied to ecological flow problems. Regional statistics have been used to predict ecologically relevant streamflow characteristics at ungauged sites to support the development of ecological response functions, with streamflow as the controlling variable [5–7]. Such statistical models depend on prior definition of the streamflow characteristics of interest and thus are of limited flexibility should other flow characteristics later emerge as important [8]. An alternative approach is the use of runoff models, which simulate an entire hydrograph for some period of interest from which any number of streamflow characteristics can subsequently be calculated [8]. Runoff models have been recommended by some authors as the tool of choice for ecological flow studies [4], while others have expressed reservations about their suitability for such applications [8,9].

There are two main criticisms related to using runoff models for application to ecological-flow studies. The first is the difficulty in transferring the calibrated model parameters from a gauged basin, where the model can be calibrated and verified, to an ungauged basin where model performance cannot be evaluated directly. This issue of predictions in ungauged catchments is an area of active research and can be addressed by different regionalization approaches [10]. However, even with perfectly estimated parameter values

(*i.e.*, the estimated parameters for an ungauged catchment correspond to what had been achieved with local model calibration) a second issue remains. This is that the models are generally calibrated on some measure of overall model performance such as the model efficiency [8,9], while biological responses to streamflow are commonly associated with specific aspects of the hydrograph, such as the long-term mean or, often more important, high- or low-flow extremes [6,11–14]. This observation raises the question: Can alternative approaches to the design and calibration of runoff models improve their ability to estimate ecologically relevant flow characteristics with a level of accuracy and precision needed to provide useful insights to the interaction between streamflow and ecosystems?

In this study, we used the HBV-light model [15–19] for runoff simulations. This model is an example of a multi-tank catchment model, with 10–15 parameters which are typically estimated by calibration. Several objective functions, each focusing on a different aspect of the hydrograph, were used to calibrate HBV-light. The aim of this study was to evaluate different objective functions for their ability to produce simulated time series that adequately preserve ecologically important flow characteristics.

2. Materials and Methods

2.1. Study Catchments

The 27 catchments used in this analysis represent parts of four Level 3 Ecoregions [20], listed east to west: Blue Ridge, Ridge and Valley, Central Appalachians, and Appalachian (Cumberland) Plateau (Figure 1). The catchments have average basin area of 829 square kilometers (km^2) (range 104–4799 km^2) and average elevation of 491 m above the North American Vertical Datum of 1988 (NAVD 88) (range 174–937 m) (Table 1). Hardwood forest and pasture are the dominant land cover in the study area. Soils are deep in the Blue Ridge ecoregion which leads to increased baseflow in comparison to the relatively thinner soils of the Appalachian Plateau and Ridge and Valley ecoregions [20] Generally, topographic slope and regolith thickness decreases from east to west, while karst development is most prominent in the Ridge and Valley [21]. Combined, these catchment characteristics produce noticeable and documented regional variations in hydrologic response and streamflow regimes [21–24].

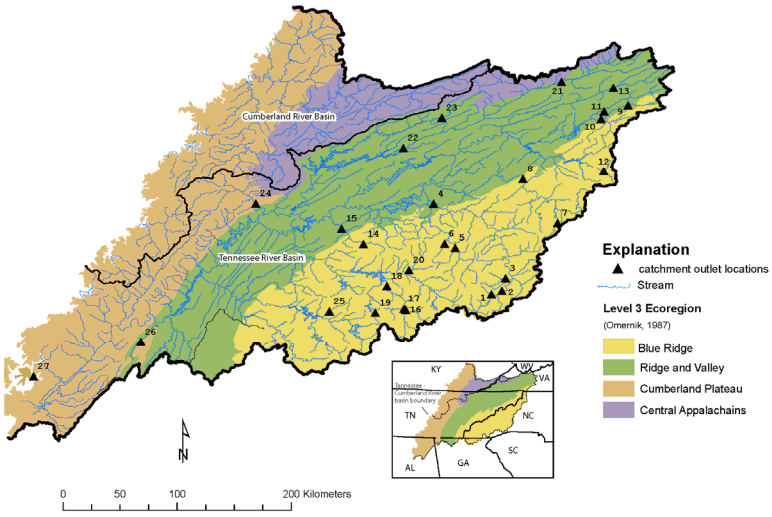


Figure 1. Catchment outlet locations for 27 basins modelled using 7 calibration schemes for HBV-light.

Table 1. U.S. Geological Survey (USGS) stream gaging sites used for model calibration and error evaluation. Latitude and longitude represent the basin outlet; ecoregion defined as the Level 3 ecoregion with the majority of the basin area; km^2, square kilometers; horizontal reference is North American Datum 1983; vertical reference is North American Vertical Datum 1988.

Map Number (Figure 1)	USGS Station Number	Latitude	Longitude	Average Elevation (m)	Primary Ecoregion (Omernik, 1987)	Basin Area (km^2)
1	03441000	35.2731	−82.7058	645	Blue Ridge	104
2	03443000	35.2992	−82.6239	628	Blue Ridge	766
3	03446000	35.3981	−82.5950	637	Blue Ridge	173
4	03455000	35.9816	−83.1611	308	Blue Ridge	4799
5	03459500	35.6350	−82.9900	712	Blue Ridge	906
6	03460000	35.6675	−83.0736	749	Blue Ridge	127
7	03463300	35.8314	−82.1842	810	Blue Ridge	112
8	03465500	36.1765	−82.4574	463	Blue Ridge	2082
9	03471500	36.7604	−81.6312	642	Blue Ridge	198
10	03473000	36.6518	−81.8440	546	Blue Ridge	785
11	03475000	36.7132	−81.8187	555	Ridge and Valley	534
12	03479000	36.2392	−81.8222	795	Blue Ridge	236
13	03488000	36.8968	−81.7462	519	Ridge and Valley	578
14	03497300	35.6645	−83.7113	337	Blue Ridge	271
15	03498500	35.7856	−83.8846	259	Blue Ridge	697
16	03500000	35.1500	−83.3797	612	Blue Ridge	361
17	03500240	35.1589	−83.3942	615	Blue Ridge	146
18	03503000	35.3364	−83.5269	537	Blue Ridge	1130
19	03504000	35.1275	−83.6186	937	Blue Ridge	135
20	03512000	35.4614	−83.3536	562	Blue Ridge	476
21	03524000	36.9448	−82.1549	457	Ridge and Valley	1382
22	03528000	36.4251	−83.3982	323	Ridge and Valley	3816
23	03531500	36.6620	−83.0949	384	Central Appalachians	828
24	03540500	35.9831	−84.5580	232	Cumberland Plateau	1815
25	03550000	35.1389	−83.9806	474	Blue Ridge	268
26	03568933	34.8975	−85.4631	202	Ridge and Valley	379
27	03574500	34.6243	−86.3064	174	Cumberland Plateau	814

Temperature and precipitation in the study area vary with longitude and elevation. Average annual temperature in the area is 13.9 degrees Celsius (°C). The warmest months of the year are July and August, and the coldest are typically January and February [25]. The Blue Ridge averages about 1350 millimeters per year (mm/y) of precipitation annually, compared to 1450 mm/y in the Cumberland Plateau and Ridge and Valley [26]. Locally, precipitation in the Blue Ridge can exceed 2000 mm/y at the highest elevations. Less than 2 percent of the precipitation comes as snow (based on 1:10 ratio of rain to snow). The streamflow regime in the study area is characterized by peak runoff typically between December and April as the result of frozen or saturated soils and low evapotranspiration rates. Summer months typically have lower streamflows because of increased temperatures and evapotranspiration rates, though occasional convective or tropical storm systems may produce locally severe flooding. Lowest flows occur in the

late-summer through the fall coinciding with continuing high temperatures and evapotranspiration rates combined with decreased precipitation (October is the driest month generally). Annual runoff for the study area varies from approximately 450 to more than 760 mm [27].

The Tennessee and Cumberland River basins (considered as one aquatic ecoregion by Abell *et al.* [28]) have the highest level of freshwater diversity in North America and possibly the most diversity for any temperate freshwater ecoregion in the world [29,30]. Included in this measure are 231 fish species (with 67 (29 percent) being endemic) along with a globally outstanding unionid mussel and crayfish fauna. Many of these species are restricted to the Tennessee and Cumberland River basins [28] (pp. 212–213). A wide range of human activities threaten these populations, including urbanization, mining, logging, agriculture, and other forms of land disturbance that alter hydrologic response [28]. In addition, the entire main channels of the Tennessee and Cumberland Rivers, together with many of their tributaries, have been impounded. Flow alteration as a result of these activities has degraded or destroyed stream habitat according to Abell *et al.* [28], with more than 57 fish species and 47 mussel species at risk in the Tennessee–Cumberland aquatic ecoregion [31] (cited in Abell *et al.* [28], p. 213).

2.2. HBV Model

The HBV model [15,16] is a simple multi-tank-type model for simulating runoff. Rainfall and air temperature data [32] as well as estimated potential evaporation data based on the American Society of Civil Engineers Penman–Monteith method [33–36] are inputs to the model, which consists of four commonly used routines: (1) snow; (2) soil moisture; (3) response; and (4) routing. These routines, or slight modifications, are commonly used in other similar models (for example PRMS; Leavesley, Lichty, Troutman, and Saindon, 1983). In the snow routine, snow accumulation and snow melt are calculated by a degree-day method [37]. The soil moisture routine represents soil–water storage, which is used in conjunction with temperature and precipitation to drive evaporation and groundwater recharge. Evaporation from the soil tank equals the potential evaporation if the relative soil moisture storage is above a certain fraction, while below that fraction a linear reduction is applied. The response routine consists of connected shallow and deep groundwater storage terms and simulates runoff by summing up three linear outflow equations representing peak, intermediate and base flow. The routing routine delivers simulated runoff to the catchment outlet based on a triangular weighting function in the routing routine.

Catchments can be separated into different elevation and vegetation zones as well as into subbasins in HBV. In this study, however, catchments were disaggregated using only different elevation zones to reduce problems of over-parameterization. Calculations were performed separately for each elevation zone according to catchment for the snow and soil-moisture routines. Groundwater storage was treated as a lumped representation for each catchment. The version of HBV used in this study, HBV-light [18], corresponds to a slightly modified version of HBV-6. HBV-light uses a warming-up period of normally one year to set state variable values according to the preceding meteorological conditions and parameter sets. A more detailed description of HBV-light can be found in [18].

2.3. Calibration

The HBV-light model was applied to the 27 catchments using a daily time step. Each catchment was separated into elevation zones of 200 m, which cover at least 5 percent of the area of their respective

catchment. Elevation zones covering less than 5 percent of the catchment area were merged with neighboring elevation zones. Rainfall and temperature data were compiled for the different elevation zones with a lapse rate of 10 percent/100 m and 0.6 °C/100 m, respectively. The long-term monthly potential evaporation data were linearly interpolated to daily values and corrected by using the deviations of the temperature to its long-term mean.

For all catchments, the first three years of input data measurements were used for the "warming-up" of the model to estimate the initial state variables. The rest of the data were divided into two equal time periods (14 years) covering the hydrological years (1 October through 30 September) from 1983 to 1996 and from 1996 to 2009. Each time period served both as calibration and validation period; when using the first time period for calibration the second time period was used for validation, and vice versa. This approach to calibration, validation, and parameterization allows us to consider distributions of parameter values derived from multiple independent realizations of the model, providing a generally robust evaluation. To address parameter uncertainty and equifinality [38], each calibration was repeated 100 times (here called calibration trials), which because of the random elements of the Genetic Algorithm and Powell optimization (GAP, [39]) used for calibration, resulted in 100 different parameterizations. The feasible parameter value ranges were defined based on previous studies (Table 2) [40].

Table 2. Parameter ranges used during the Genetic Algorithm and Powell optimization (GAP) calibrations within HBV-light. (°C, degrees Celsius; mm, millimeter; D, day).

Parameter	Explanation	Minimum	Maximum	Unit
Snow Routine				
TT	Threshold temperature	−2	2.5	°C
CFMAX	Degree-day factor	0.5	10	$mm \cdot °C^{-1} \cdot D^{-1}$
SFCF	Snowfall correction factor	0.5	1.2	-
CFR	Refreezing coefficient	0	0.1	-
CWH	Water holding capacity	0	0.2	-
Soil Routine				
FC	Maximum storage in soil box	100	550	mm
LP	Threshold for reduction of evaporation (relative storage in the soil box)	0.3	1	-
BETA	Shape coefficient	1	5	-
Response Routine				
PERC	Maximal flow from upper to lower box	0	4	$mm \cdot D^{-1}$
UZL	Maximal storage in the soil upper zone	0	70	mm
K0	Recession coefficient (upper box, upper outflow)	0.1	0.5	D^{-1}
K1	Recession coefficient (upper box, lower outflow)	0.01	0.2	D^{-1}
K2	Recession coefficient (lower box)	0.00005	0.1	D^{-1}
Routing Routine				
MAXBAS	Routing, length of weighting function	1	5	D

We considered seven different objective functions for calibration, which consisted of either single or combined statistical criteria evaluating the fit between observed and simulated values (Tables 3 and 4) to

assess the influence of an objective function on the value of the simulated ecological indicators. The objective functions were chosen to represent different statistical aspects of streamflow. The combinations of criteria were defined to evaluate different aspects simultaneously; for example, combination 2 (C2) included Reff, MARE, Spearman, and Volume Error (see Table 3 for a description of the criteria). Reff and MARE are sensitive to peaks and low flows, respectively, and therefore help evaluate performance with respect to extreme discharge values. Volume Error expresses how well the model predicts overall runoff volume for the simulation period, whereas the Spearman rank coefficient reflects the model's success in replicating the overall timing and magnitude of discharge. Each objective function was used to calibrate the model for each time period, resulting in 14 simulated time series (seven objective functions for two different calibration periods) of streamflow for each catchment modeled.

Table 3. Definitions criteria used in objective functions for the automatic calibration trials using the Genetic Algorithm and Powell optimization (GAP) algorithm.

Criterion	Description	Definition
Reff	Model efficiency	$1 - \frac{\sum(Q_{\text{obs}} - Q_{\text{sim}})^2}{\sum(Q_{\text{obs}} - \overline{Q_{\text{obs}}})^2}$
LogReff	Efficiency for log(Q)	$1 - \frac{\sum(\ln Q_{\text{obs}} - \ln Q_{\text{sim}})^2}{\sum(\ln Q_{\text{obs}} - \ln \overline{Q_{\text{obs}}})^2}$
Lindström	Lindström measure	$Reff - 0.1\frac{\lvert\sum(Q_{\text{obs}} - Q_{\text{sim}})\rvert}{\sum(Q_{\text{obs}})}$
MARE	Measure based on the Mean Absolute Relative Error [(1)]	$1 - \frac{1}{n}\sum\frac{\lvert Q_{\text{obs}} - Q_{\text{sim}}\rvert}{Q_{\text{obs}}}$
Spearman	Spearman rank correlation [(2)]	$\frac{\sum(R_{\text{obs}} - \overline{R_{\text{obs}}})(S_{\text{sim}} - \overline{S_{\text{sim}}})}{\sqrt{\sum(R_{\text{obs}} - \overline{R_{\text{obs}}})^2}\sqrt{\sum(S_{\text{sim}} - \overline{S_{\text{sim}}})^2}}$
VolumeError	Volume error	$1 - \frac{\lvert\sum(Q_{\text{obs}} - Q_{\text{sim}})\rvert}{\sum(Q_{\text{obs}})}$

[(1)] Where n is the number of days; [(2)] Where R_{obs} and S_{sim} are the ranks of Q_{obs} and Q_{sim}, respectively.

Table 4. The three combination objective functions used during the Genetic Algorithm and Powell optimization (GAP) calibrations within HBV-light. The criteria were weighted equally in each case. See Table 3 for a more detailed specification of each of the criteria.

Combined Objective Function	Criteria
C1	Reff, LogReff, VolumeError
C2	Reff, MARE, Spearman, VolumeError
C3	Spearman, VolumeError

2.4. Evaluation

The choice of the SFCs is based on studies of Knight *et al.* [6], which identified 12 specific streamflow characteristics, from a larger suite identified in Knight *et al.* [41], as most appropriate indicators for fish species richness in the study area (Table 5). All SFCs were computed using the simulated runoff of each catchment that was calibrated with one of the seven objective functions and for the two different calibration and validation time periods. The value of each streamflow characteristic was determined for

both time periods based on the measurement data. All indices were computed using the free EflowStats R-Package [42].

Table 5. Definition of streamflow characteristics used in this study (adapted and modified from Knight *et al.*, 2014 and Thomson and Archfield, 2014) (mm/day, millimeters per day; -, no units; %, percent).

Streamflow Characteristic	Abbreviation	Description	Units
Magnitude			
Mean annual runoff	MA41	Annual mean daily streamflow	mm/day
Maximum October runoff	MH10	Mean maximum October streamflow across the period of record	mm/day
Lowest 15% of daily runoff	Flowperc	85% exceedance of daily mean streamflow for the period of record	mm/day
Rate of streamflow recession	RA7	Median change in log of streamflow for days in which the change is negative across the period of record	mm/day
Ratio			
Average 30-day maximum runoff	DH13	Mean annual maximum of a 30-day moving average streamflow divided by the median for the entire record	–
Stability of runoff	TA1	Measure of the constancy of a flow regime by dividing daily flows into predetermined flow classes	–
Frequency			
Frequency of moderate floods	FH6	Average number of high-flow events per year that are equal to or greater than three times the median annual flow for the period of record	number/year
Frequency of moderate floods	FH7	Average number of high-flow events per year that are equal to or greater than three times the median annual flow for the period of record	number/year
Variability			
Variability of March runoff	MA26	Standard deviation for March streamflow divided by the mean streamflow for March	–
Variability in high-flow pulse duration	DH16	100 times the standard deviation for the yearly average high-flow pulse durations (daily flow greater than the 75th percentile) divided by the mean of the yearly average high pulse durations	%
Variability of low-flow pulse count	FL2	100 times the standard deviation for the average number of yearly low-flow pulses (daily flow less than the 25th percentile) divided by the mean low-flow pulse counts	%
Date			
Timing of annual minimum runoff	TL1	Julian date of annual minimum flow occurrence	Julian day

For each objective function, 100 calibration trials were accomplished per catchment for both periods (1983–1996 and 1996–2009), producing 100 independently optimized parameter sets per catchment per simulation period. For each objective function and streamflow characteristic, the sources of uncertainty in the results were analyzed. The spread reflects both differences in behavior among the 27 catchments and uncertainty among the parameter sets, but the relative importance of these two sources of variability is not uniform. The variability because of differences between catchments was analyzed by computing the medians of the streamflow characteristics over the 100 runs per catchment. To be able to compare the median values, normalization was carried out by dividing the median values by the corresponding observed flow characteristic value. For analyzing the spread resulting from parameter uncertainty, the ranges over 100 runs per catchment were divided by the range over the median values of the different

catchments. The spread because of parameter uncertainty was compared to the variation between the different catchments.

To quantify the performance of objective functions in representing the different flow characteristics, Spearman rank correlation coefficients and Nash-Sutcliffe efficiencies (NSEs) were computed between the (median) simulated and observed flow characteristic values of the 27 different catchments. Where NSE of 1.0 corresponds to identical flow characteristic values between simulated and observed runoff time series for each catchment, a Spearman rank correlation coefficient of 1.0 only requires the order of observed and simulated flow characteristic values to be the same.

3. Results

The model efficiencies that could be achieved for the different catchments varied from 0.64 to 0.91 (calibration) and 0.61 to 0.90 (validation), indicating reasonably good runoff simulation with the calibrated HBV-light model. As an example of the performance of the simulations with regard to the streamflow characteristics, the results for two indices (DH16 (variability in high-flow pulse duration) and MA41 (mean annual runoff)) for one catchment (03455000) are shown in Figure 2. Each plot contains 28 boxplots (one for each combination of an objective function, time period and calibration or validation). Each of the boxplots is based on 100 streamflow characteristic values obtained by using the 100 different parameter sets per catchment for the simulations. In both cases, there were clear deviations of the flow characteristics computed from the simulated time series compared to the observed runoff series as indicated by the red lines (red line represents observed SFC value). The streamflow characteristic DH16 was largely underestimated, especially for period 1 (1983–1996) (Figure 2a). The spread among the 100 different simulations was considerably larger for period 2 (1996–2009) than for period 1. For SFCs such as MA41 (Figure 2b), the performance differences in predicting the streamflow characteristic were prominent between the four combinations of calibration and validation periods.

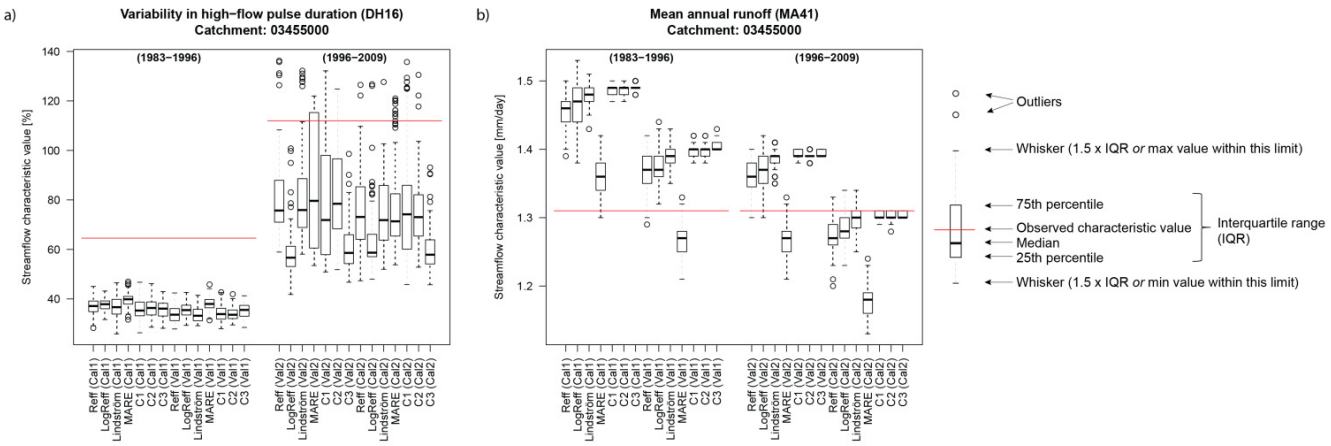


Figure 2. Boxplots for catchment 4 (03455000) and (**a**) streamflow characteristic DH16 (Variability in high-flow pulse duration); (**b**) streamflow characteristic MA41 (Mean annual runoff). Cal1 and Cal2 are calibration of period 1, respectively period 2, whereas Val1 and Val2 are validation of period 1, respectively period 2.

The agreement between observed and simulated flow characteristics varied considerably among the different catchments (Figure 3). Each plot contains 28 boxplots (one for each combination of an objective function, time period and calibration or validation). Each boxplot is based on 27 values (one value per catchment), which were normalized by dividing the median streamflow characteristic value based on simulated runoff by the corresponding streamflow characteristic value computed based on the observed runoff time series. The spread between the different catchments is much smaller for the streamflow characteristic MA41 (mean annual runoff) than for the other flow characteristics. Except for the criteria LogReff and MARE, MA41 was reproduced well for both calibration periods, whereas values were slightly underestimated when being validated on period 1 and slightly overestimated when validated on period 2. Both MA41 (mean annual runoff) and MH10 (maximum October runoff) were reproduced less well for parameter sets derived by calibration based on the criteria LogReff and MARE, both of which are more sensitive to low flow conditions than the other criteria.

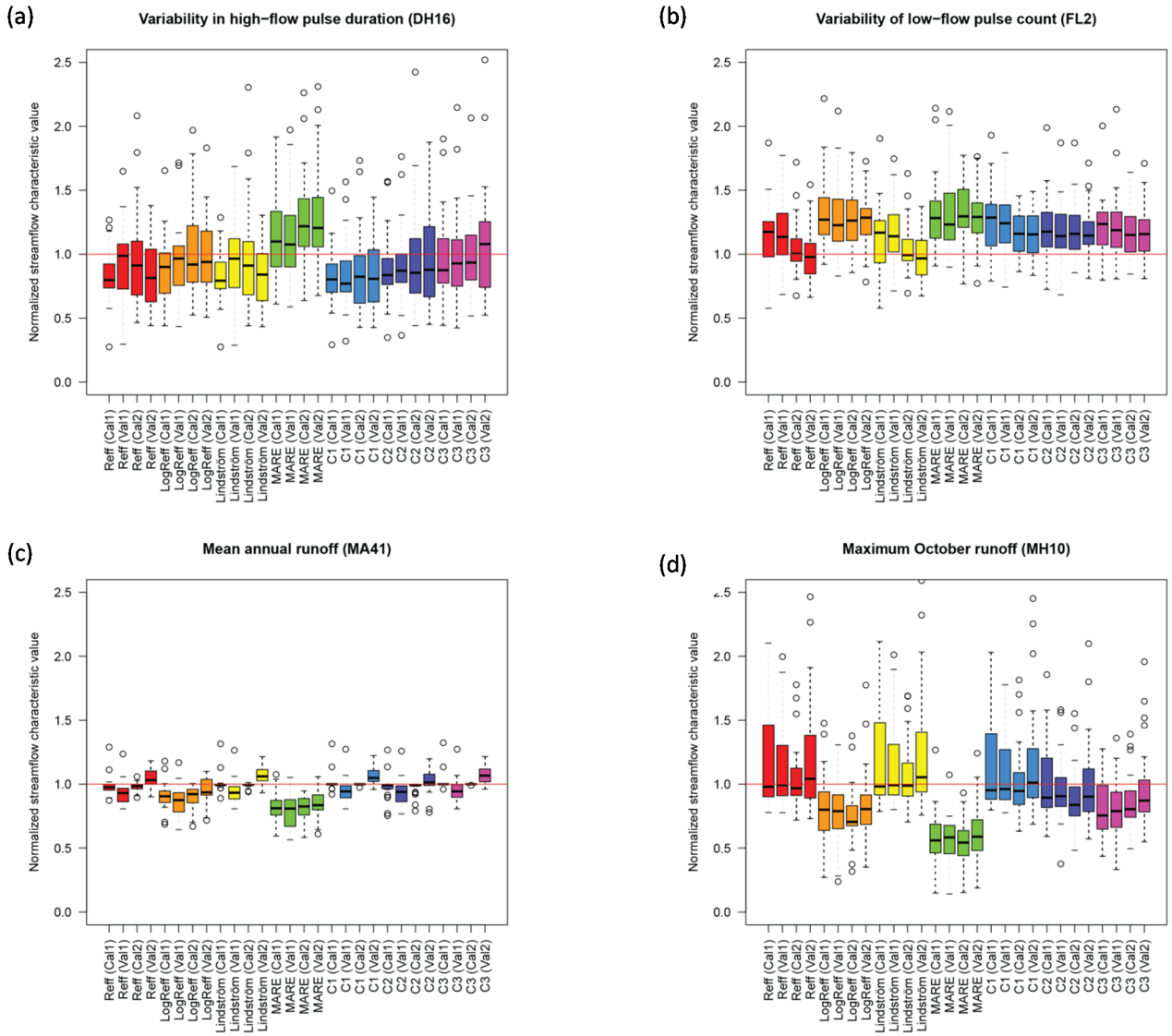


Figure 3. *Cont.*

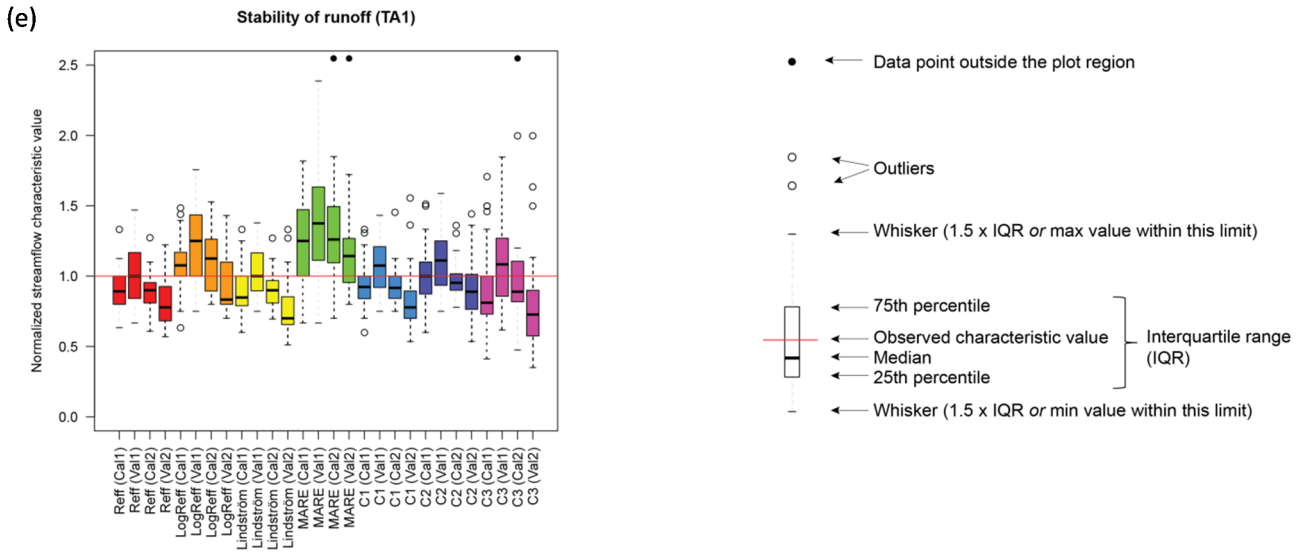


Figure 3. Normalized median flow characteristic values for five different flow characteristics: (**a**) DH16 (Variability in high-flow pulse duration); (**b**) FL2 (Variability of low-flow pulse count); (**c**) MA41 (Mean annual runoff); (**d**) MH10 (Maximum October runoff) and (**e**) TA1 (Stability of runoff). Each color corresponds to an objective function. Per objective function, the four boxplots represent (from left to right) calibration period 1 (Cal1), validation period 1 (Val1), calibration period 2 (Cal2) and validation period 2 (Val2). Each boxplot is based on 27 normalized median flow characteristic values, one value for each of the 27 catchments. Medians were computed over 100 runs per catchment. Normalization was carried out by dividing the median values by the corresponding observed flow characteristic value.

The distribution of the 27 relative ranges (per catchment—Dividing the range over the 100 runs per catchment by the range over the 27 median catchment values) is a measure for the consistency over the different catchments (Figure 4). While for some cases there was a low variation (indicated by narrow distributions of relative range), for many cases a considerable variation was observed. For calibrations based on the Nash-Sutcliffe efficiency, for instance, the median relative range varied from around 0.1 for MA41 (mean annual runoff) to above 1 for FL2 (variability of low-flow pulse count).

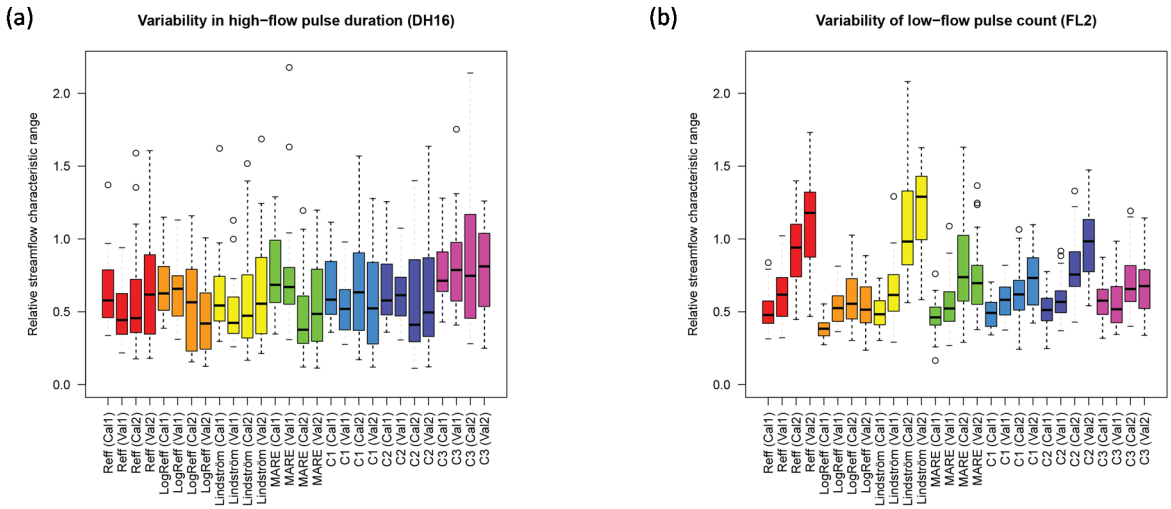


Figure 4. *Cont.*

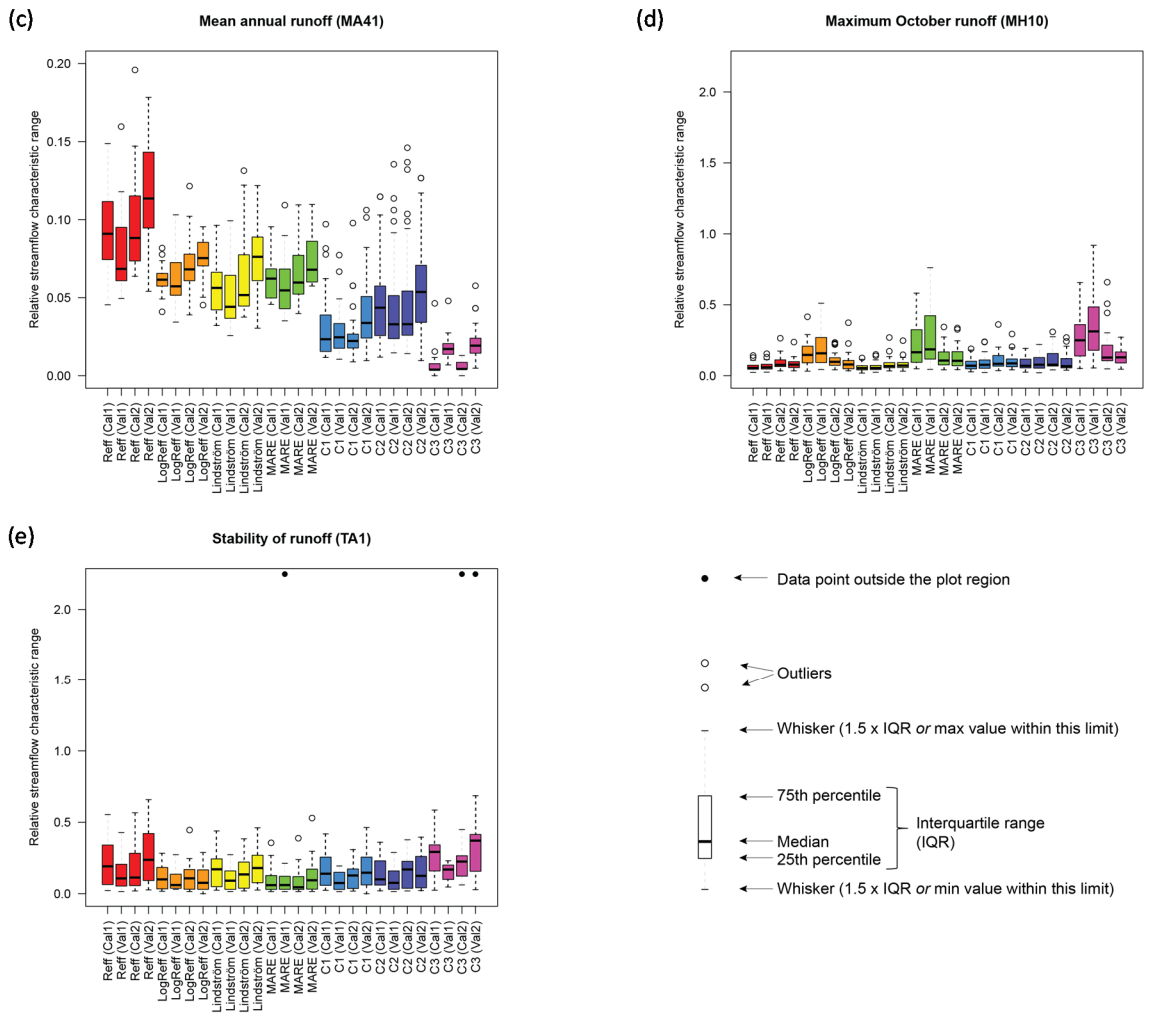


Figure 4. Relative ranges as a measure for parameter uncertainty for streamflow characteristics (**a**) DH16 (Variability in high-flow pulse duration); (**b**) FL2 (Variability of low-flow pulse count); (**c**) MA41 (Mean annual runoff); (**d**) MH10 (Maximum October runoff) and (**e**) TA1 (Stability of runoff). Each color corresponds to an objective function. Per objective function, the four boxplots represent (from left to right) calibration period 1 (Cal1), validation period 1 (Val1), calibration period 2 (Cal2) and validation period 2 (Val2). Each boxplot is based on 27 values, one value for each of the 27 catchments. Relative ranges were computed by dividing the range over the 100 runs per catchment by the range over the 27 median catchment values. Note that the Mean annual runoff (MA41) has been plotted on a different scale.

Agreement among the different streamflow characteristics and the different objective functions varied considerably (Figure 5). Comparison of streamflow characteristics based on observed runoff series against the medians of those obtained from simulated time series allows evaluating the agreement in relation to the variation between catchments. These scatter plots show that the agreement varied considerably among both the different streamflow characteristics and the different objective functions. While only plots with flow characteristics calculated for the first calibration period are shown, results were similar for the other calibration and validation periods. The performance for all streamflow characteristics and all combinations of calibration/validation periods were evaluated using the Spearman rank correlation coefficients (Table 6), which evaluates how well the relative ranking of the indices between the catchments is captured, and

the model efficiencies (Table 7), which evaluate how well the exact values were predicted. Typically, the values were similar for periods 1 and 2, when the parameterizations obtained by calibration for the respective period were used, resulting in a median difference of 0.015 for the Spearman Rank correlation and 0.0855 for NSE. In general, results are expected to be poorer for the validation period in comparison to the calibration period; however, for the respective validation periods the values were only slightly lower (median difference of −0.0215 (Spearman) and −0.029 (NSE)). This indicates that results were similar for the two periods and were similar when looking at the validation periods. The average median percent error for estimated streamflow characteristics was almost always less than zero, indicating that the objective functions used for model calibration typically underestimated each of the 12 streamflow characteristics being evaluated (Table 8).

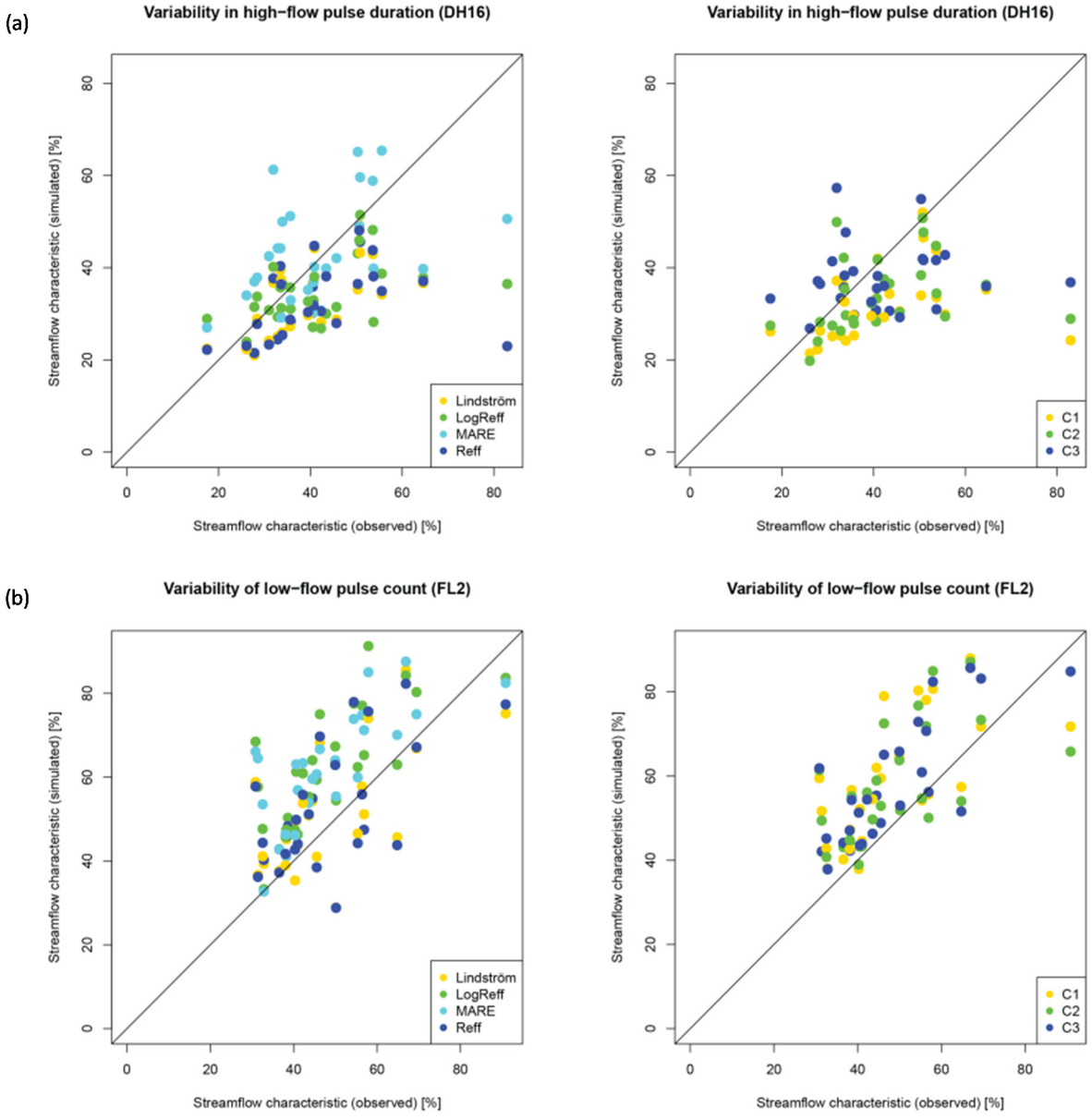


Figure 5. *Cont.*

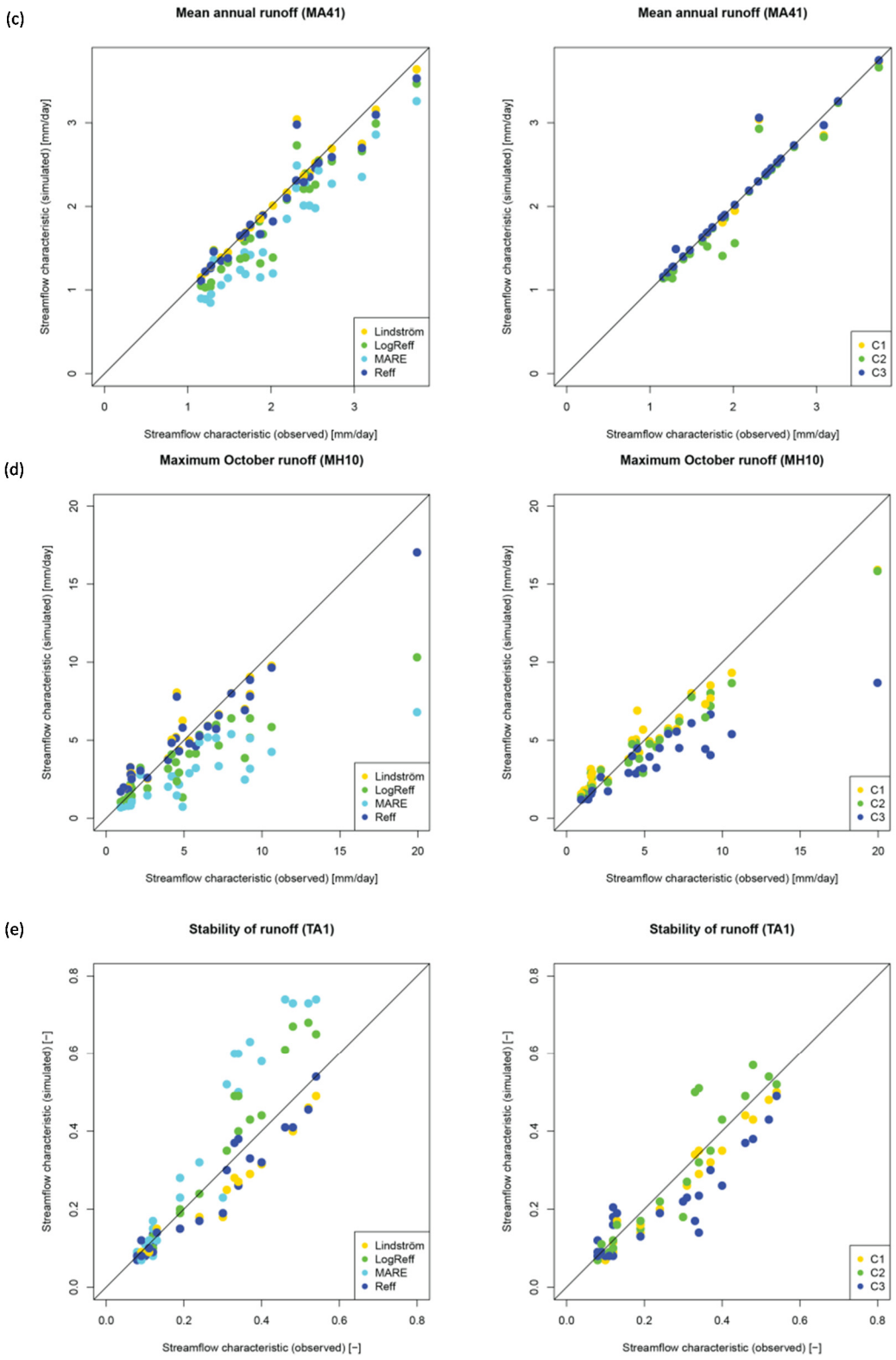


Figure 5. Scatterplots for the streamflow characteristics (**a**) DH16 (Variability in high-flow pulse duration); (**b**) FL2 (Variability of low-flow pulse count); (**c**) MA41 (Mean annual runoff); (**d**) MH10 (Maximum October runoff) and (**e**) TA1 (Stability of runoff) for calibration period 1. The points represent the median value of all 100 calibration trials in each catchment based on single criteria objective functions (**left column**) and multi-criteria objective functions (**right column**).

Table 6. Spearman rank correlation coefficients between objective functions (horizontal) and streamflow characteristics (vertical) based on observed respective simulated streamflow (for each group of four values: upper − left = calibration period 1 (Cal1), upper − right = validation period 2 (Val2), lower − left = validation period 1 (Val1), lower − right = calibration period 2 (Cal2)). Colors are ranging from white (for a Spearman rank correlation of 0) to dark green (for a Spearman rank correlation of 1).

	Reff		LogReff		Lindström		MARE		C1		C2		C3	
MA41	0.973	0.978	0.930	0.927	0.980	0.983	0.919	0.918	0.980	0.981	0.947	0.928	0.981	0.986
	0.957	0.991	0.929	0.947	0.961	0.998	0.926	0.950	0.961	1.000	0.952	0.979	0.962	1.000
MH10	0.930	0.831	0.874	0.853	0.916	0.837	0.834	0.829	0.941	0.837	0.958	0.874	0.918	0.898
	0.960	0.940	0.862	0.868	0.958	0.934	0.822	0.829	0.957	0.918	0.942	0.903	0.885	0.933
Flowperc	0.796	0.978	0.810	0.986	0.790	0.961	0.814	0.979	0.808	0.980	0.810	0.983	0.685	0.867
	0.778	0.985	0.808	0.996	0.781	0.980	0.804	0.996	0.803	0.995	0.806	0.996	0.683	0.897
RA7	0.736	0.724	0.877	0.885	0.726	0.735	0.888	0.896	0.870	0.873	0.851	0.892	0.696	0.797
	0.756	0.836	0.930	0.930	0.719	0.775	0.848	0.902	0.878	0.919	0.880	0.917	0.744	0.789
DH13	0.977	0.938	0.974	0.948	0.971	0.908	0.960	0.960	0.981	0.945	0.976	0.945	0.926	0.691
	0.955	0.866	0.976	0.937	0.955	0.877	0.964	0.957	0.971	0.910	0.978	0.885	0.871	0.573
TA1	0.972	0.929	0.968	0.943	0.977	0.906	0.947	0.974	0.968	0.884	0.960	0.899	0.875	0.766
	0.936	0.956	0.933	0.966	0.952	0.942	0.884	0.936	0.958	0.948	0.942	0.964	0.904	0.924
FH6	0.943	0.851	0.916	0.906	0.935	0.875	0.728	0.863	0.953	0.916	0.900	0.921	0.569	0.663
	0.926	0.888	0.853	0.931	0.931	0.898	0.634	0.855	0.942	0.930	0.901	0.919	0.498	0.613
FH7	0.948	0.933	0.881	0.889	0.949	0.935	0.810	0.887	0.967	0.945	0.965	0.952	0.688	0.563
	0.927	0.951	0.842	0.889	0.941	0.960	0.763	0.805	0.945	0.967	0.944	0.967	0.480	0.520
MA26	0.849	0.917	0.789	0.906	0.855	0.920	0.704	0.858	0.894	0.923	0.903	0.915	0.631	0.856
	0.752	0.932	0.699	0.894	0.782	0.935	0.672	0.829	0.821	0.933	0.831	0.928	0.381	0.769
DH16	0.534	0.645	0.443	0.662	0.503	0.673	0.402	0.471	0.510	0.745	0.525	0.683	0.145	0.482
	0.429	0.549	0.421	0.654	0.410	0.514	0.346	0.645	0.526	0.659	0.511	0.650	0.094	0.518
FL2	0.521	0.443	0.740	0.628	0.609	0.449	0.734	0.703	0.709	0.602	0.684	0.668	0.755	0.594
	0.548	0.617	0.659	0.604	0.579	0.659	0.641	0.626	0.672	0.711	0.620	0.695	0.616	0.628
TL1	0.477	0.394	0.643	0.520	0.471	0.347	0.612	0.753	0.603	0.330	0.531	0.428	0.574	0.418
	0.407	0.112	0.646	0.546	0.418	0.065	0.623	0.777	0.497	0.362	0.531	0.201	0.600	0.280

Table 7. Nash-Sutcliffe efficiencies between objective functions (horizontal) and streamflow characteristics (vertical) based on observed respective simulated streamflow (for each group of four values: upper − left = calibration period 1 (Cal1), upper − right = validation period 2 (Val2), lower − left = validation period 1 (Val1), lower − right = calibration period 2 (Cal2)). Colors are ranging from white (for Nash-Sutcliffe efficiencies of 0 or lower) to dark green (for a Nash-Sutcliffe efficiency of 1).

	Reff		LogReff		Lindström		MARE		C1		C2		C3	
MA41	0.917	0.936	0.840	0.881	0.936	0.933	0.584	0.626	0.946	0.939	0.922	0.927	0.949	0.930
	0.858	0.967	0.746	0.835	0.900	0.993	0.490	0.554	0.914	0.999	0.875	0.965	0.916	1.000
MH10	0.848	0.820	−0.627	0.570	0.841	0.796	−3.942	−1.220	0.820	0.871	0.796	0.879	−1.630	0.663
	0.859	0.934	−0.931	0.332	0.874	0.926	−5.692	−2.258	0.848	0.926	0.756	0.850	−1.367	0.667

Table 7. *Cont.*

	Reff		LogReff		Lindström		MARE		C1		C2		C3	
Flowperc	0.416	0.749	0.611	0.837	0.356	0.660	0.647	0.960	0.463	0.680	0.614	0.804	0.170	0.477
	0.484	0.868	0.569	0.967	0.491	0.820	0.465	0.966	0.538	0.848	0.591	0.939	0.373	0.669
RA7	0.209	0.281	0.071	0.193	0.279	0.370	−0.420	−0.284	−0.043	−0.063	−0.229	−0.197	−9.226	−7.224
	−0.628	−0.230	0.369	0.385	−0.608	−0.277	0.156	0.186	0.276	0.252	0.190	0.231	−5.173	−4.088
DH13	0.372	−0.164	0.884	0.472	−0.601	−1.895	0.910	0.858	0.797	0.522	0.770	0.874	−7.603	−20.044
	0.638	0.427	0.919	0.748	0.437	−0.030	0.814	0.914	0.902	0.813	0.672	0.817	−4.235	−14.891
TA1	0.898	0.432	0.856	0.882	0.829	0.108	0.672	0.803	0.918	0.477	0.886	0.749	0.502	-1.020
	0.863	0.912	0.718	0.845	0.892	0.926	0.548	0.685	0.881	0.974	0.839	0.953	0.806	0.705
FH6	0.709	0.628	−1.354	−0.967	0.660	0.559	−7.331	−4.461	0.513	0.502	0.210	0.282	−3.781	−5.629
	0.714	0.622	−0.788	−0.465	0.717	0.612	−4.768	−3.426	0.736	0.680	0.533	0.522	−2.536	−4.020
FH7	0.746	0.756	−0.440	−1.246	0.585	0.600	−0.752	−1.837	0.769	0.725	0.842	0.820	−13.413	−22.837
	0.813	0.826	0.290	−0.242	0.801	0.820	−0.260	−0.612	0.912	0.930	0.932	0.954	−9.425	−11.728
MA26	0.618	0.849	0.080	0.033	0.582	0.832	−0.418	−1.114	0.789	0.882	0.848	0.872	−4.116	−4.256
	0.331	0.862	0.184	0.320	0.324	0.886	0.178	−0.513	0.500	0.894	0.564	0.878	−1.898	−2.343
DH16	−3.044	−0.329	−3.375	0.050	−3.323	−0.307	−0.463	−0.371	−3.727	−0.006	−2.768	0.192	−3.474	−0.562
	−0.937	−0.182	−2.056	0.186	−1.012	−0.234	−1.025	0.006	−1.535	−0.092	−1.562	0.119	−2.785	−0.309
FL2	0.118	−1.176	−0.469	−1.557	0.201	−0.931	−0.556	−1.448	−0.266	−0.827	−0.167	−1.773	0.139	−0.948
	−0.040	−1.198	−0.530	−1.841	0.056	−1.123	−0.759	−1.703	−0.203	−0.409	−0.132	−1.246	-0.104	−1.018
TL1	−0.376	−4.676	−0.211	−3.016	−0.310	−5.502	−0.361	−2.672	−0.017	−4.483	−0.196	−4.053	−0.023	−2.708
	−0.505	−4.322	−0.250	−3.892	−0.518	−4.338	−0.557	−2.218	−0.400	−4.503	−0.489	−5.932	0.021	−3.529

Table 8. Median percent error for streamflow characteristics by model objective function for calibration period 1 (Cal1).

Objective Function	MA41	MH10	RA7	TA1	DH13	FH7	FH6	FL2	MA26	DH16	TL1	E85	Average Median Error (Percent)
Lindström	−0.6	−1.8	−25.0	−15.2	−18.1	−23.0	−12.0	16.8	9.1	−20.8	3.7	19.1	−5.6
LogReff	−9.5	−20.0	−50.0	7.7	−9.5	−37.5	−27.0	26.9	−7.3	−10.0	4.8	15.2	−9.7
MARE	−18.9	−44.0	−57.1	25.0	−7.4	−44.4	−41.4	28.2	−19.6	9.9	5.5	−7.3	−14.3
Reff	−2.5	−2.1	−18.2	−10.8	−14.7	−20.0	−12.0	17.5	9.8	−20.2	4.2	9.8	−4.9
C1	0.0	−4.8	−50.0	−7.7	−13.1	−19.0	−14.1	28.6	4.9	−19.7	3.4	29.9	−5.1
C2	−0.8	−10.6	−42.9	0.0	−7.5	−14.0	−18.2	17.7	2.2	−16.4	4.0	13.2	−6.1
C3	0.0	−24.5	−44.4	−18.9	−18.9	−69.3	−37.6	23.6	−28.1	−12.5	3.4	24.1	−16.9
Average Median Percent Error	−4.6	−15.4	−41.1	−2.8	−12.7	−32.5	−23.2	22.8	−4.2	−12.8	4.1	14.9	–

4. Discussion

In the absence of observed data, environmental flow studies necessarily rely on some form of streamflow estimation to model the response of aquatic ecology to alteration of the streamflow regime. Knight *et al.* [23] and Murphy *et al.* [8] raised the question of validity and began evaluation of model accuracies for predicting known ecologically-relevant streamflow characteristics. Murphy *et al.* [8] and Shrestha *et al.* [9] highlight that typical calibration approaches, often focused on daily, monthly, or annual mean values, are inadequate when predicting more subtle aspects of the flow regime. An increasing body of work is making use of statistical modeling approaches to address hydrologic and hydro-ecological

questions [5,7,43–45]. However, as already stated by Murphy *et al.* [8] and Shrestha *et al.* [9], runoff models have advantages as well as limitations, particularly in regard to developing streamflow time series reflecting land cover, human population, or climatic projections. As such, runoff models should be closely evaluated to better understand if the calibration approaches and predictive accuracies yield results amenable to their end use.

While the HBV-light model was used in this study, there is little reason to assume that results would be discernibly different if another calibrated runoff model were used. Partly this reflects the fact that most mechanistic runoff models are fundamentally similar in concept and application, using more or less the same or similar routines. Fundamentally, if calibration is used, the simulated series are fitted to the observed series according to some objective function, and regardless of the specific model being used, this fit does not ensure agreement in all possible aspects of the hydrograph shape.

The accuracy of prediction and appropriateness of calibration is important in the context of environmental flow application as error of predicting flow-regime components will be translated and probably amplified as error in estimating ecological response. A given approach to model calibration will lead to accurate prediction of the runoff with regard to the used objective function measure, however accurate prediction of other aspects may be lacking. For example, Knight *et al.* [41] (Figure 2) published linear functions representing the 80th quantile upper-bound relationship of specialized insectivore scores to three streamflow characteristics (TA1, FH6, and RA7; see Table 5 for definitions). Following Murphy *et al.* [8], we use these relations to evaluate the accuracy of streamflow characteristic predictions as well as predicted ecological response based on the seven calibration approaches discussed herein for a single model (catchment 03488000). Using the equations from Knight *et al.* [41] and simulated streamflow presented in this paper, values of insectivore scores varied from 0.49 to 0.87 for RA7, 0.53 to 0.8 for TA1, and 0.58 to 0.84 for FH6 (Table 9; Figure 6). While median percent difference error for estimated specialized insectivore score for RA7 was a modest 8.2 percent under the estimate using observed data, individual departures from the observed values ranged from −19.7 to 42.6 percent for RA7, −13.1 to 31.1 percent for TA1, and −10.8 to 29.2 percent for FH6. Model results in this example are similar to those for a regional regression model reported by Murphy *et al.* [8] (9 percent difference for streamflow characteristic and 16 percent over estimation for insectivore score using HBV-light. Results presented here are considerably different than those for a rainfall-runoff model example from Murphy *et al.* [8], showing 90 percent overestimated for the same ecological score.

The objective functions used for model calibration resulted overall in an underprediction of the 12 streamflow characteristics being evaluated (Table 8). The general underprediction of the flow characteristics is a result similar to that seen in Murphy *et al.* [8] where a TOPMODEL application calibrated on mean annual flow was evaluated in the context of predicting the same streamflow characteristics. The median errors presented here are within plus-or-minus 30 percent of observed values, proposed by Kennard *et al.* [46] as an acceptable band of uncertainty, for 8 to 12 streamflow characteristics (out of 12) depending on the objective function (Figure 7, Table 8). This is in stark contrast to the rainfall runoff model evaluated in Murphy *et al.* [8]) where 13 of 19 streamflow characteristics were outside this band. While similar patterns are seen in overall model results, the calibration approaches evaluated in this paper appear to have provided more accurate estimates across the flow regime as defined by these characteristics. These results can be attributed both to the use of 100 parameter sets, which resulted in more robust flow characteristic estimations, and the use of different objective functions. Parameter uncertainty

was substantial for many streamflow characteristics depending on which objective function was used. Despite this, high model efficiencies could still be achieved in many cases when using the median of 100 calibration trials as a more robust prediction for streamflow characteristics.

Table 9. Comparison of selected streamflow characteristics based on simulated and observed streamflow time series for a single model location (site 13 (03488000)) and calibration period 1 (Cal1). (TA1, RA7, and FH6, defined in Table 5; values in parentheses represent the specialized insectivore score using the associated streamflow characteristic value based on linear equations presented in Knight *et al.* [41], Figure 2; hydro, percent error for streamflow characteristic derived from simulated and observed streamflow time series; eco, percent error for specialized insectivore score based on streamflow characteristic derived from simulated and observed streamflow time series).

Objective Function (see Table 3 for Definitions)	RA7		Percent Error	TA1		Percent Error	FH6		Percent Error
	Simulated	Observed	Hydro/Eco	Simulated	Observed	Hydro/Eco	Simulated	Observed	Hydro/Eco
Lindström	0.14 (0.49)		27.3/−19.7	0.4 (0.55)		−16.7/−9.8	13 (0.59)		13.4/−9.2
LogReff	0.1 (0.66)		−9.1/8.2	0.67 (0.75)		39.6/23	10.08 (0.7)		−12/7.7
MARE	0.06 (0.83)		−45.5/36.1	0.73 (0.8)		52.1/31.1	6.62 (0.84)		−42.2/29.2
Reff	0.125 (0.55)		13.6/−9.8	0.41 (0.56)		−14.6/−8.2	13.38 (0.58)		16.8/−10.8
C1	0.12 (0.57)	0.11 (0.61)	9.1/−6.6	0.43 (0.57)	0.48 (0.61)	−10.4/−6.6	12.92 (0.59)	11.46 (0.65)	12.7/−9.2
C2	0.09 (0.7)		−18.2/14.8	0.57 (0.68)		18.8/11.5	12.38 (0.62)		8/−4.6
C3	0.05 (0.87)		−54.5/42.6	0.38 (0.53)		−20.8/−13.1	6.54 (0.84)		−42.9/29.2

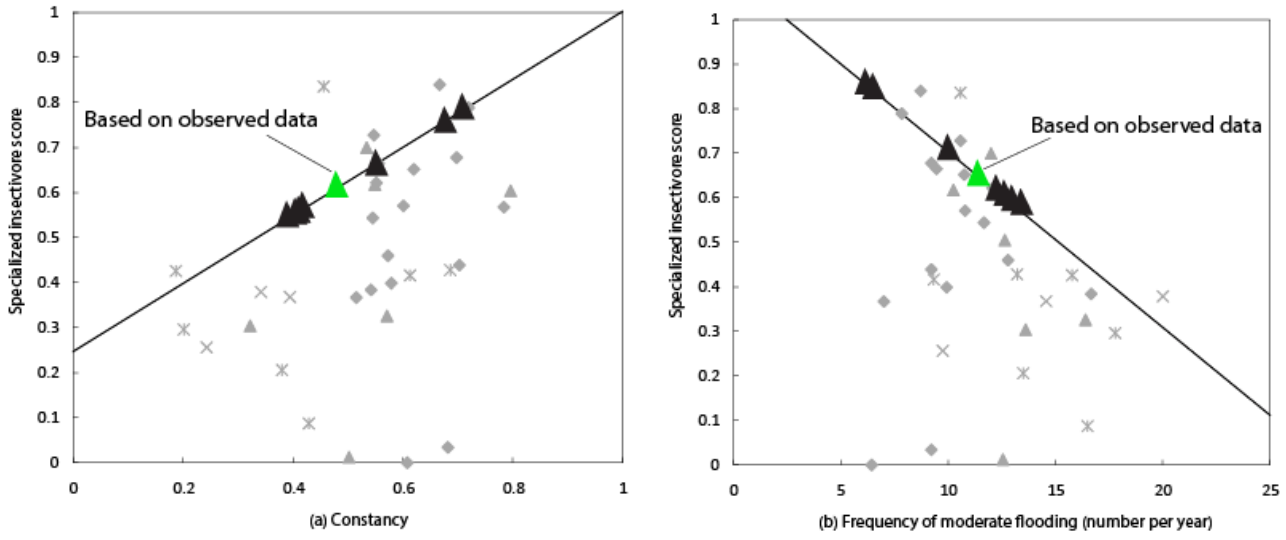


Figure 6. *Cont.*

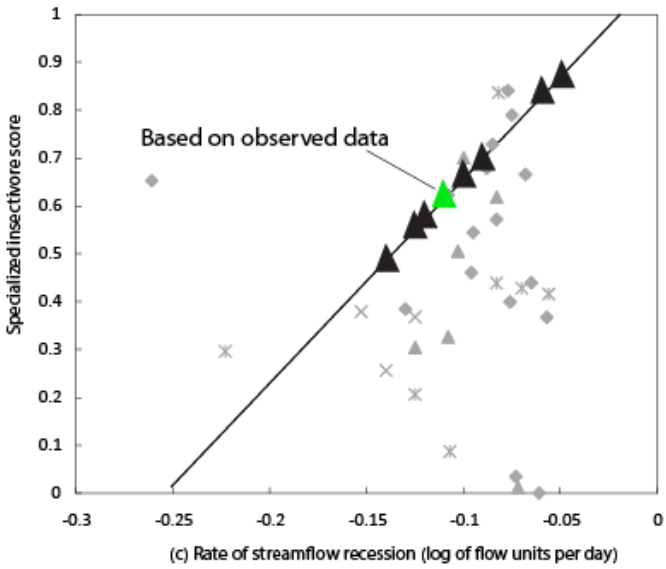


Figure 6. Example of an ecological flow application by comparison of estimated values for three streamflow characteristics for site 13 (03488000) (Table 1, Figure 1) and calibration period 1 (Cal1). (**a**) Constancy; (**b**) Frequency of moderate flooding (number per year) and (**c**) Rate of streamflow recession (log of flow units per day). Black triangles represent model estimated values based on the seven objective functions. Green triangle represents streamflow characteristics based on observed data. Values for RA7 (Rate of streamflow recession) were multiplied by negative 1 to convert values to those in the original analysis. Thin black lines represent 80th percentile quantile regression lines based on the 33 data point (grayed) in the background used by Knight *et al.* [41]. (Figure modified from Knight *et al.* [41]).

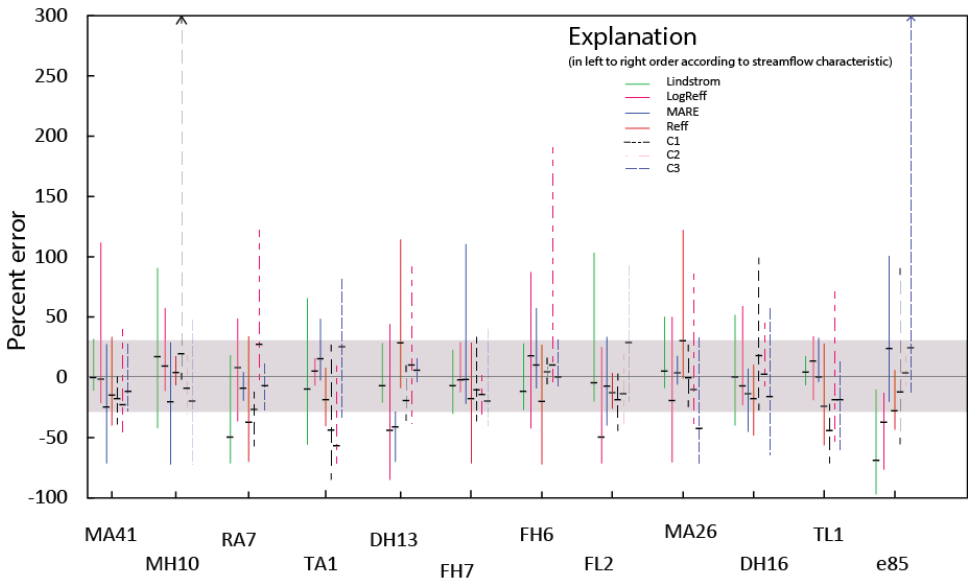


Figure 7. Minimum, maximum, and median percent errors according to objective function and streamflow characteristic for calibration period 1 (Cal1). Each vertical bar is based on the median error for the 27 catchments. The gray band in the center of the figure represents ±30 percent difference [46] Vertical bars with arrows indicate the maximum percent error exceeded the axis scale.

While the low average median percentage error would indicate a good performance with regard to the estimated flow characteristics, the scatter plots and computed Nash-Sutcliffe efficiencies and Spearman rank correlations reveal a slightly different picture. Spearman rank correlations were rather high for many of the objective functions and streamflow characteristics. For many of those objective function and flow characteristic combinations, however, Nash-Sutcliffe efficiencies were much lower. This shows that, although a clear bias might be observed in the predicted streamflow characteristic values, the order between the catchments was preserved quite well. In practice it might be more important to determine how well the flow characteristics are reproduced relative to the variation among catchments in the region than to determine the relative error value. When evaluating the scatter plots (Figure 5), low values of the Nash-Sutcliffe efficiencies indicated that the represented variability was relatively low, and the low Spearman rank correlations indicated that some flow characteristics that were not similar on a ranking scale were estimated correctly for the different catchments.

Considering individual streamflow characteristics, a pattern in predictive accuracy is evident. Most notably, streamflow characteristics that reflect average conditions (MA41, MA26, TA1, and TL1) were predicted quite well, with average median percent errors ranging from 2.8 to 4.6 percent absolute (Table 8). However, for some of these characteristics, especially TL1, the relative variation of the simulated values among the catchments were rather poor (Tables 6 and 7). Aspects of the hydrograph representative of high-flow conditions (MH10, FH7, FH6, DH13, DH16, and RA7) were underpredicted consistently (between 12.7 and 41.1 percent), with individual model calibrations underpredicting values up to 70 percent under observed. Low-flow characteristics were overpredicted (FL2 and E85) by 22.8 and 14.9 percent respectively. This appears to indicate that the model, regardless of calibration, may be retaining water during high-flow periods and allowing it to release during low-flow periods. The considerable underprediction of RA7 (rate of streamflow recession) indicates that higher flow events receded at a slower rate, which is suggestive of water stored in groundwater, and subsequently abundant groundwater discharge. The underprediction of RA7 and overprediction of low-flow characteristics are complementary.

MA41 (mean annual runoff) was predicted extremely well, particularly when using those calibrations where the objective function included the volume error as criterion, which is expected as this criterion is equivalent to the mean annual runoff. Predictions of MA41 also performed quite well when calibrated using the Nash-Sutcliffe efficiency. This performance might be attributed to the sensitivity of the Nash-Sutcliffe efficiency for high flows, which could reduce the error in the estimation of mean annual runoff. As noted by Murphy *et al.* [8], inclusion of ecological flow characteristics as criteria in calibrations may yield better simulations.

5. Conclusions

The accuracy of simulated runoff resulting from seven objective functions was evaluated in this paper by comparing streamflow characteristics based on observed and predicted streamflow time series. While the ultimate goal is to produce the most accurate simulated streamflow time series at ungauged catchments based on the transfer of calibrated parameter sets from gauged to ungauged catchments, the comparison in this study addresses an important part of the total uncertainty, namely the uncertainty related to the prediction accuracy specific streamflow characteristics that were not part of the calibration routine. The primary conclusion is that good model performance in terms of objective functions, such as the frequently

used Nash-Sutcliffe model efficiency, does not ensure that all flow characteristics computed from these simulations will correspond to those derived from observed runoff. This is an important consideration that is often overlooked by users of model output who use simulated time series for various analyses, supporting resource allocation decisions, or establishing flow policy. While expecting simulated runoff series to agree with the observed in all possible aspects is unreasonable, this analysis serves as a further reminder of the substantial errors possible, using ecological flow characteristics as the example.

Two novel approaches were used in this study. First, we evaluated the effectiveness of seven objective functions for simulating streamflow time series and subsequent streamflow characteristic calculations. This allowed for critical examination of the importance of the objective function choice, as results differed substantially among objective functions. Results indicate there was no single best calibration strategy, but not surprisingly, different strategies provided better predictions for different streamflow characteristics. However, there was some indication that the combined objective functions, which evaluate the runoff simulations in different aspects, might be generally more suitable across a range of flow characteristics. Second, parameter uncertainty was explicitly considered by using the combination of 100 different equally possible parameter sets for each calibration trial instead of the typical single optimal calibrated parameter set. Our results confirmed the value of this approach by showing that different parameter sets can be similar with respect to the objective function used (similarity between the Nash-Sutcliffe for example) but differ greatly with respect to other characteristics. We demonstrated that using only one parameter set could result in substantial uncertainties, which can be reduced by using the values based on several parameter sets as more robust estimation.

More research is needed to determine which objective functions are most useful to ensure acceptable simulations of ecological flow characteristics, or other regime-defining characteristics. One suitable approach beyond the objective functions used in this paper might be to include streamflow characteristics of particular interest as objective functions in the calibration. This corresponds to the suggestion to include various hydrological signatures as diagnostic tools [47]. The fact that simulation-based flow characteristics varied largely depending upon which objective functions were used indicates that there is a considerable potential to improve model calibrations by considering specific flow characteristics when evaluating model performance during calibration. While it can be expected that performances improve when a certain streamflow characteristic is explicitly included in the objective function, it is less clear which criteria should be included to ensure acceptable simulations for calculation of streamflow characteristics in general. Further research is therefore motivated to explore which criteria to include in the objective function to obtain streamflow simulations that preserve as many streamflow characteristics as possible.

Acknowledgments

This paper is a product of discussions and activities that took place at the U.S. Geological Survey John Wesley Powell Center for Analysis and Synthesis as part of the workgroup focusing on Water Availability for Ungauged Rivers (https://powellcenter.usgs.gov/). Funding for this research was provided by the Tennessee Wildlife Resources Agency, the National Park Service, the U.S. Geological Survey Cooperative Water Program, and the University of Zurich. The use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government.

Author Contributions

Rodney Knight and Jan Seibert conceived the initial ideas for this study; Marc Vis performed the simulations; Jan Seibert, Marc Vis, Sandra Pool and Rodney Knight analyzed the results; all authors contributed to writing of the paper.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Rantz, S.E. *Stream Hydrology Related to the Optimum Discharge for King Salmon Spawning in the Northern California Coast Ranges*; U.S. Geological Survey Water-Supply Paper 1779-AA: Washington, DC, USA, 1964; p. 15.
2. Tennant, D.L. Instream Flow Regimens for Fish, Wildlife, Recreation and Related Environmental Resources. *Fisheries* **1976**, *1*, 6–10.
3. Olden, J.D.; Poff, N.L. Redundancy and the choice of hydrologic indices for characterizing streamflow regimes. *River Res. Appl.* **2003**, *19*, 101–121.
4. Poff, N.L.; Richter, B.D.; Arthington, A.H.; Bunn, S.E.; Naiman, R.J.; Kendy, E.; Acreman, M.; Apse, C.; Bledsoe, B.P.; Freeman, M.C.; *et al.* The ecological limits of hydrologic alteration (ELOHA): A new framework for developing regional environmental flow standards. *Freshw. Biol.* **2010**, *55*, 147–170.
5. Carlisle, D.M.; Wolock, D.M.; Meador, M.R. Alteration of streamflow magnitudes and potential ecological consequences: A multiregional assessment. *Front. Ecol. Environ.* **2011**, *9*, 264–270.
6. Knight, R.R.; Murphy, J.C.; Wolfe, W.J.; Saylor, C.F.; Wales, A.K. Ecological limit functions relating fish community response to hydrologic departures of the ecological flow regime in the Tennessee River basin, United States. *Ecohydrology* **2014**, *7*, 1262–1280.
7. Sanborn, S.C.; Bledsoe, B.P. Predicting streamflow regime metrics for ungauged streamsin Colorado, Washington, and Oregon. *J. Hydrol.* **2006**, *325*, 241–261.
8. Murphy, J.C.; Knight, R.R.; Wolfe, W.J.; Gain, W.S. Predicting Ecological Flow Regime at Ungaged Sites: A Comparison of Methods. *River Res. Appl.* **2013**, *29*, 660–669.
9. Shrestha, R.R.; Peters, D.L.; Schnorbus, M.A. Evaluating the ability of a hydrologic model to replicate hydro-ecologically relevant indicators. *Hydrol. Process.* **2014**, *28*, 4294–4310.
10. Hrachowitz, M.; Savenije, H.H.G.; Blöschl, G.; McDonnell, J.J.; Sivapalan, M.; Pomeroy, J.W.; Arheimer, B.; Blume, T.; Clark, M.P.; Ehret, U.; *et al.* A decade of Predictions in Ungauged Basins (PUB)—A review. *Hydrol. Sci. J.* **2013**, *58*, 1198–1255.
11. Clausen, B.; Biggs, B. Relationships between benthic biota and hydrological indices in New Zealand streams. *Freshw. Biol.* **1997**, *38*, 327–342.
12. Clausen, B.; Biggs, B.J. Flow variables for ecological studies in temperate streams: Groupings based on covariance. *J. Hydrol.* **2000**, *237*, 184–197.
13. Poff, N.L.; Ward, J.V. Implications of Streamflow Variability and Predictability for Lotic Community Structure: A Regional Analysis of Streamflow Patterns. *Can. J. Fish. Aquat. Sci.* **1989**, *46*, 1805–1818.
14. Puckridge, J.T.; Walker, K.F.; Costelloe, J.F. Hydrological persistence and the ecology of dryland rivers. *Regul. Rivers Res. Manag.* **2000**, *16*, 385–402.

15. Bergström, S. *Development and Application of a Conceptual Runoff Model for Scandinavian Catchments*; SMHI: Norrköping, Sweden, 1976; No. RHO 7, p. 134.
16. Bergström, S. *The HBV Model: Its Structure and Applications*; SMHI Hydrology: Norrköping, Sweden, 1992; p. 35.
17. Lindström, G.; Johansson, B.; Persson, M.; Gardelin, M.; Bergström, S. Development and test of the distributed HBV-96 hydrological model. *J. Hydrol.* **1997**, *201*, 272–288.
18. Seibert, J.; Vis, M.J.P. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrol. Earth Syst. Sci.* **2012**, *16*, 3315–3325.
19. Singh, V.P. *Computer Models of Watershed Hydrology*; Water Resources Publications: Highlands Ranch, CO, USA, 1995.
20. Omernik, J.M. Ecoregions of the Conterminous United States. *Ann. Assoc. Am. Geogr.* **1987**, *77*, 118–125.
21. Wolfe, W.; Haugh, C.; Webbers, A.; Diehl, T. Preliminary Conceptual Models of the Occurrence, Fate, and Transport of Chlorinated Solvents in Karst Regions of Tennessee. Department of Interior, US Geological Survey, Water Resources Investigations Report 97-4097. Available online: http://pubs.usgs.gov/wri/wri974097/new4097.pdf (accessed on 3 April 2015).
22. Hoos, A.B. Recharge Rates and Aquifer Hydraulic Characteristics for Selected Drainage Basins in Middle and East Tennessee. Department of the Interior, US Geological Survey, Water Resources Investigations Report 90-4015, 34. Available online: http://pubs.water.usgs.gov/wri904015/ (accessed on 24 June 2010).
23. Knight, R.R.; Gain, W.S.; Wolfe, W.J. Modelling ecological flow regime: An example from the Tennessee and Cumberland River basins. *Ecohydrology* **2012**, *5*, 613–627.
24. Law, G.S.; Tasker, G.D.; Ladd, D.E. Streamflow-Characteristic Estimation Methods for Unregulated Streams of Tennessee. US Geological Survey, Scientific Investigations Report 2009–5159, 212. p, 1 Plate. Available online: http://pubs.usgs.gov/sir/2009/5159/ (accessed on 16 June 2010).
25. U.S. Department of Commerce Climatography of the United States No. 85 Divisional Normals and Standard Deviations of Temperature, Precipitation, and Heating and Cooling Degree Days 1971–2000 (And Previous Normals Periods) Section 1: Temperature. United States Department of Commerce, Washington, DC, USA, 2007.
26. U.S. Department of Commerce Climatography of the United States No. 85 Divisional Normals and Standard Deviations of Temperature, Precipitation, and Heating and Cooling Degree Days 1971–2000 (And Previous Normals Periods) Section 2: Precipitation. United States Department of Commerce, Washington, DC, USA, 2007.
27. Moody, D.W.; Chase, E.B.; Aronson, D.A. *National Water Summary 1985—Hydrologic Events and Surface-Water Resources*; United States Geological Survey Water-Supply Paper 2300: Chapter on Tennessee Surface-Water Resources; United States Geological Survey, Reston, VA, USA, 1986; pp. 425–429.
28. Abell, R.A.; Olson, D.M.; Dinerstein, E.; Hurley, P.T.; Diggs, J.T.; Eichbaum, W.; Walters, S.; Wettengel, W.; Allnutt, T.; Loucks, C.J.; *et al. Freshwater Ecoregions of North America: A Conservation Assessment.*; Island Press: Washington, DC, USA, 2000.
29. Olson, D.M.; Dinerstein, E. The Global 200: A Representation Approach to Conserving the Earth's Most Biologically Valuable Ecoregions. *Conserv. Biol.* **1998**, *12*, 502–515.
30. Etnier, D.A.; Starnes, W.C. *The fishes of Tennessee*; The University of Tennessee Press: Knoxville, TN, USA, 1993.

31. Master, L.L.; Flack, S.R.; Stein, B.A.; Conservancy, N. *Rivers of Life: Critical Watersheds for Protecting Freshwater Biodiversity*; Nature Conservancy: Arlington, VA, USA, 1998.
32. Thornton, P.E.; Thornton, M.M.; Mayer, B.W.; Wilhelmi, N.; Wei, Y.; Devarakonda, R.; Cook, R.B. *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 2*; Oak Ridge National Laboratory: Oak Ridge, TN, USA, 2014.
33. Monteith, J.L. Evaporation and Environment. In *The State and Movement of Water in Living Organism 19th Symposia of the Society Experimental Biology*; University Press: Cambridge, UK, 1965; pp. 205–234.
34. Walter, I.; Allen, R.; Elliott, R.; Itenfisu, D.; Brown, P.; Jensen, M.; Mecham, B.; Howell, T.; Snyder, R.; Eching, S.; *et al.* The ASCE Standardized Reference Evapotranspiration Equation. PREPARED BY Task Committee on Standardization of Reference Evapotranspiration of the Environmental and Water Resources Institute. Available online: http://kimberly.uidaho.edu/water/asceewri/ascestzdetmain2005.pdf (accessed on 3 April 2015).
35. Rotstayn, L.D.; Roderick, M.L.; Farquhar, G.D. A simple pan-evaporation model for analysis of climate simulations: Evaluation over Australia. *Geophys. Res. Lett.* **2006**, *33*, L17715.
36. Hobbins, M.; Wood, A.; Streubel, D.; Werner, K. What Drives the Variability of Evaporative Demand across the Conterminous United States? *J. Hydrometeorol.* **2012**, *13*, 1195–1214.
37. Rango, A.; Martinec, J. Revisiting the degree-day method for snowmelt computations. *J. Am. Water Resour. Assoc.* **1995**, *31*, 657–669.
38. Beven, K.; Freer, J. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.* **2001**, *249*, 11–29.
39. Seibert, J. Multi-Criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrol. Earth Syst. Sci.* **2000**, *4*, 215–224.
40. Seibert, J. Regionalisation of parameters for a conceptual rainfall-runoff model. *Agric. For. Meteorol.* **1999**, *98–99*, 279–293.
41. Knight, R.R.; Brian Gregory, M.; Wales, A.K. Relating streamflow characteristics to specialized insectivores in the Tennessee River Valley: A regional approach. *Ecohydrology* **2008**, *1*, 394–407.
42. Thompson, J.; Archfield, S. *The EflowStats R package Introduction to EflowStats*; United States Geological Survey: Reston, VA, USA, 2014.
43. Castellarin, A.; Camorani, G.; Brath, A. Predicting annual and long-term flow-duration curves in ungauged basins. *Adv. Water Resour.* **2007**, *30*, 937–953.
44. McManamay, R.A. Quantifying and generalizing hydrologic responses to dam regulation using a statistical modeling approach. *J. Hydrol.* **2014**, *519*, 1278–1296.
45. Zhu, Y.; Day, R.L. Regression modeling of streamflow, baseflow, and runoff using geographic information systems. *J. Environ. Manag.* **2009**, *90*, 946–953.
46. Kennard, M.J.; Mackay, S.J.; Pusey, B.J.; Olden, J.D.; Marsh, N. Quantifying uncertainty in estimation of hydrologic metrics for ecohydrological studies. *River Res. Appl.* **2010**, *26*, 137–156.
47. Gupta, H.V.; Wagener, T.; Liu, Y. Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrol. Process.* **2008**, *22*, 3802–3813.

Paper II

Hydrol. Earth Syst. Sci., 21, 5443–5457, 2017
https://doi.org/10.5194/hess-21-5443-2017

Hydrology and Earth System Sciences
Open Access
EGU

Streamflow characteristics from modeled runoff time series – importance of calibration criteria selection

Sandra Pool[1], **Marc J. P. Vis**[1], **Rodney R. Knight**[2], **and Jan Seibert**[1,3,4]

[1]Department of Geography, University of Zurich, Zurich, Switzerland
[2]U.S. Geological Survey Lower Mississippi-Gulf Water Science Center, 640 Grassmere Park, Suite 100, Nashville, TN 37211, USA
[3]Department of Earth Sciences, Uppsala University, Uppsala, Sweden
[4]Department of Physical Geography, Stockholm University, Stockholm, Sweden

Correspondence to: Sandra Pool (sandra.pool@geo.uzh.ch)

Received: 17 October 2016 – Discussion started: 19 October 2016
Revised: 7 July 2017 – Accepted: 15 September 2017 – Published: 6 November 2017

Abstract. Ecologically relevant streamflow characteristics (SFCs) of ungauged catchments are often estimated from simulated runoff of hydrologic models that were originally calibrated on gauged catchments. However, SFC estimates of the gauged donor catchments and subsequently the ungauged catchments can be substantially uncertain when models are calibrated using traditional approaches based on optimization of statistical performance metrics (e.g., Nash–Sutcliffe model efficiency). An improved calibration strategy for gauged catchments is therefore crucial to help reduce the uncertainties of estimated SFCs for ungauged catchments. The aim of this study was to improve SFC estimates from modeled runoff time series in gauged catchments by explicitly including one or several SFCs in the calibration process. Different types of objective functions were defined consisting of the Nash–Sutcliffe model efficiency, single SFCs, or combinations thereof. We calibrated a bucket-type runoff model (HBV – Hydrologiska Byråns Vattenavdelning – model) for 25 catchments in the Tennessee River basin and evaluated the proposed calibration approach on 13 ecologically relevant SFCs representing major flow regime components and different flow conditions. While the model generally tended to underestimate the tested SFCs related to mean and high-flow conditions, SFCs related to low flow were generally overestimated. The highest estimation accuracies were achieved by a SFC-specific model calibration. Estimates of SFCs not included in the calibration process were of similar quality when comparing a multi-SFC calibration approach to a traditional model efficiency calibration. For practical applications, this implies that SFCs should preferably be estimated from targeted runoff model calibration, and modeled estimates need to be carefully interpreted.

1 Introduction

Reliable runoff information is fundamental for many water resources-related tasks such as flood prevention, drought mitigation, management of drinking water supply and hydropower, or river restoration. Runoff modeling is a tool that can be used to create runoff time series when observed time series are not available. Runoff simulations usually focus on either representing the general shape of the hydrograph or on accurately simulating specific streamflow characteristics relevant to a respective application. However, the extraction of streamflow characteristics (SFCs) from a simulated time series may produce poor estimates when these characteristics were not included in model calibration. Ecologically relevant SFCs are properties of the annual streamflow hydrograph defining the structure and functioning of aquatic and riparian biodiversity (Richter et al., 1996; Poff et al., 1997). The accurate prediction of streamflow characteristics is a core determinate to defining how streamflow and aquatic communities relate. A large number of SFCs have been suggested to characterize ecologically relevant aspects of the flow regime (Tharme, 2003) and have become the basis for decision-support systems integrating resource management with ecological response (Cartwright et al., 2017).

Published by Copernicus Publications on behalf of the European Geosciences Union.

Multivariate regression or runoff models are used to estimate SFCs when observed streamflow time series data are not available (Hailegeorgis and Alfredsen, 2016). The estimation of SFCs with linear regression usually relates a single SFC to catchment characteristics such as climate, land cover, and geographic and geologic variables (e.g., Sanborn and Bledsoe, 2006; Carlisle et al., 2010; Knight et al., 2012). This approach is inflexible in a sense that the regression is SFC-specific and does not allow for analysis of potential water-use and land management (Murphy et al., 2013). These disadvantages can be partially overcome by applying runoff models. Simulated streamflow time series from runoff models can be used to calculate any SFC and, by changing model input and parameters, different scenarios such as climate change, groundwater withdrawals, land use, and riverine change can be simulated (Poff et al., 2010; Murphy et al., 2013; Olsen et al., 2013; Shrestha et al., 2014). While statistical models such as multiple linear regressions often provide greater accuracy (Murphy et al., 2013), runoff models provide opportunities for also evaluating climate or land-use change scenarios.

Runoff models are used in both ecohydrology and hydrological modeling as tools to simulate specific aspects of the runoff regime. The terms, SFCs or ecological flow indices, are often used to refer to such specific aspects of the flow regime in ecohydrology studies, whereas the more recently introduced term, hydrological signatures, has been used in hydrological modeling (Jothityangkoon et al., 2001; Wagener et al, 2007). Hydrological signatures can often support a physical interpretation of the way a catchment functions and are seen as valuable metrics especially for modeling ungauged catchments (Jothityangkoon et al., 2001), for selecting appropriate model structures (Euser et al., 2013) or guiding model parameter selection in a meaningful way (Yilmaz et al., 2008), and for classifying catchments (Wagener et al., 2007; Sawicz et al., 2011). Regardless of the terminology and the ultimate goal, the basic goal is the quantification of certain aspects of a streamflow time series. In this paper, we use the term SFC as equivalent to hydrological signature, but generally prefer the term SFC to emphasize their ecological relevance.

Estimated streamflow characteristics are prone to significant errors when calculated from simulated time series (Murphy et al., 2013; Shrestha et al., 2014; Vis et al., 2015). This is due in part to the objective functions used for evaluating the model error such as the commonly used model efficiency (Nash and Sutcliffe, 1970) or volume error, which do not ensure that a model reproduces particular streamflow characteristics. These objective functions subsequently guide model parameter calibration, which strongly influences the simulated hydrograph (for an overview, see Pfannerstill et al., 2014) in terms of annual, seasonal, and monthly volumes and magnitudes. For example, Vis et al. (2015) compared model simulation from calibrations based on only the model efficiency with calibrations based on the combination of multiple objectives such as model efficiency, model efficiency of log-transformed flow, volume error, and Spearman rank correlation. All these calibration approaches tended to overestimate low flows and underestimate medium and high-flow-related SFCs. Estimation accuracy varied greatly between SFCs, with absolute biases between 3 and 33 %. Large differences in estimation accuracy are also reported by Shrestha et al. (2014) and Ryo et al. (2015). Their multi-objective calibration approach resulted in runoff simulations favoring high flows at the expense of the estimation accuracy of low flows. The large variability in estimated SFC accuracy as well as the bias in the estimates can generally be observed independently of the model used to simulate the runoff time series (Caldwell et al., 2015). A remedy to this large variability and bias is to incorporate SFCs into model calibration schemes. For example, Westerberg et al. (2011) and Pfannerstill et al. (2014) focused on specific evaluation points or segments of the flow-duration curve (FDC) during model calibration. Both studies report better overall performance for the simulated hydrograph with a FDC-based calibration compared to a more traditional calibration approach using, for example, the model efficiency (Nash and Sutcliffe, 1970). However, runoff models calibrated using FDC have to be constrained by additional SFCs if one is interested in the exact timing of events or when snow-related runoff processes are of importance (Westerberg et al., 2011). Yilmaz et al. (2008) combined information on different segments of the FDC with the runoff ratio and the rainfall–runoff lag time to guide model parameter selection in terms of primary catchment functions. These hydrologically meaningful signatures generally improved hydrograph simulation, but their value was limited for the process of vertical redistribution of excess rainfall in the catchment. In a recent study, Kiesel et al. (2017) compared estimates of ecologically relevant SFCs simulated from model calibrations using different objective functions including SFCs and the Kling–Gupta efficiency (Gupta et al., 2009). They found that including all SFCs of interest in the model calibration resulted in better SFC estimates than a calibration using the Kling–Gupta efficiency. Instead of aiming at a well-simulated, general hydrograph, Hingray et al. (2010) and Olsen et al. (2013) focused on certain aspects of the streamflow regime that were considered most important. Their results, which are echoed by Murphy et al. (2013), suggest that the runoff model performs reasonably well for the aspects on which it is calibrated, whereas it only modestly represents other runoff characteristics. Hence, developing an approach to increase the accuracy of estimated SFCs from runoff model time series continues to be an open challenge in hydrological modeling.

This study expands on the study of Vis et al. (2015) where various combinations of traditionally used objective functions were evaluated with respect to a suite of ecologically relevant SFCs. Their model calibrations with the model efficiency (R_{eff}) outperformed multi-objective model calibrations (different combinations of R_{eff}, log-transformed flow, volume error, and Spearman rank correlation) for the investi-

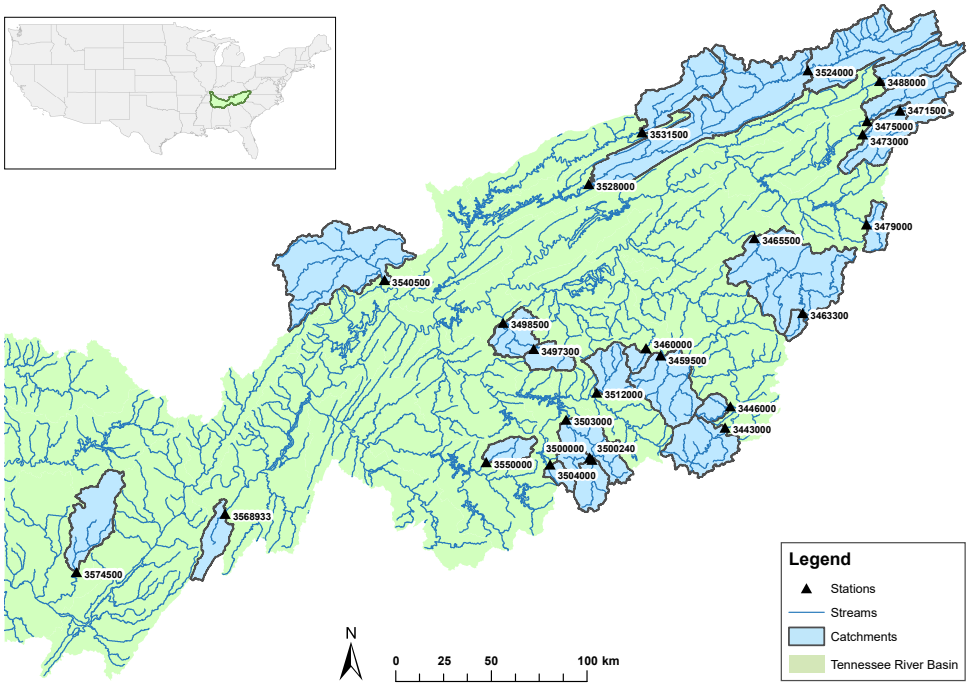


Figure 1. Location of the 25 study catchments in the Tennessee River basin (Table 1 in Vis et al., 2015, for more information).

gated SFCs. It was furthermore hypothesized that the explicit consideration of SFCs in runoff model calibration could reduce bias in estimated SFCs. The main objective of this study was therefore to assess the potential for a runoff model calibrated using specific aspects of the flow regime to more accurately estimate a suite of SFCs as compared to using a model efficiency-based calibration approach. The general approach was based on the idea that most information essential for estimating SFCs is preserved in the simulated hydrograph by including selected SFCs in model calibration. Our modeling approach relies on catchments with observed runoff time series and therefore does not answer the question of how to simulate SFCs in ungauged or altered catchments. However, the prediction of runoff for ungauged catchments benefits from an improved and informed calibration strategy for gauged catchments, which is used in the subsequent regionalization. For regionalization approaches we refer to studies such as Yadav et al. (2007), Viglione et al. (2013), or Westerberg et al. (2016).

The following questions are addressed in this paper:

1. How well is a single SFC simulated when that SFC is used as the model objective function? (Objective function is the SFC of interest.)

2. How well is a single SFC simulated when the model objective function contains one or multiple other SFCs? (Objective function can include the SFC of interest, but generally contains one or multiple other SFCs.)

3. How does the accuracy of estimated SFCs vary between traditional calibration approaches and those where the SFCs of interest are included? (Objective functions are different combinations of SFC(s) and the model efficiency.)

Throughout this study, we refer to traditional and "SFC-based" objective functions. Traditional objective functions were defined as efficiency criteria based on statistical performance metrics computed from (transformed) model residuals (e.g., R_{eff} or volume error). In contrast, "SFC-based" objective functions evaluate specific hydrograph aspects, such as event frequencies, timing, or variability of runoff, that are of ecological relevance in our study region.

2 Materials and methods

2.1 Catchment locations and characteristics

The study catchments are all located in the 106 000 km^2 Tennessee River basin in the southeastern United States (Fig. 1), which is one of the most diverse temperate freshwater ecosystems in the world (Abell et al., 2000). A large number of endemic fish species and a unique assemblage of mussels, crayfish, and salamanders make the Tennessee River basin an excellent area for ecohydrological studies (Abell et al., 2000). From a study published by Knight et al. (2008), 25 catchments in the Tennessee River basin with observed streamflow time series (U.S. Geological Survey, 2016b), pre-

cipitation (U.S. Department of Commerce, 2007a), temperature (U.S. Department of Commerce, 2007b), and potential evaporation data (Rotstayn et al., 2006) were selected. The catchment areas range between 100 and 4800 km^2 with elevations ranging from 174 to 937 m (U.S. Geological Survey, 2016a). Land cover for the study catchments is predominantly hardwood forest and pasture. Air temperature and precipitation vary between catchments according to both catchment elevation and longitude. Mean annual air temperature in the 25 catchments varies between 9.3 and 14.7 °C, and annual precipitation varies from 1500 to 2020 mm, with fall being slightly drier and less than 8 % of annual precipitation falling as snow. Runoff is highest in winter and lowest in summer, ranging from 400 to 1300 mm a^{-1} (millimeters per year). Variability in soil thickness (Omernik, 1987), regolith thickness, karst development, and topographic slope (Hoos, 1990; Wolfe et al., 1997; Law et al., 2009) are documented as asserting the most influence on runoff.

2.2 Selection of SFCs

Thirteen SFCs assessed in this study were chosen for use in model scenarios based on discernible functional connections with fish community diversity (Knight et al., 2008, 2014). This set of 13 SFCs represents each of the major flow regime components commonly used in ecological studies (e.g., Olden and Poff, 2003; Arthington et al., 2006; Caldwell et al., 2015): magnitude, ratio, frequency, variability, and date (Table 1). For this study the SFCs were additionally grouped according to flow conditions (mean, low, and high flow), because different aspects of the hydrograph have been shown to be sensitive to the objective function used for model calibration (for an overview, see Pfannerstill et al., 2014). The SFCs were calculated using the U.S. Geological Survey (2014) EflowStats R package. Please note that some of the tested SFCs (DH13, ML20, MA26, DH16, and FL2) are defined as scaled with the median, mean, or total runoff. The scaling leads to SFC values that are dependent on flow magnitudes. The magnitude of the simulation error for DH13, ML29, MA26, DH16, and FL2 is therefore dependent on runoff magnitudes, whereas the sign of the simulation error is not affected by the normalization.

2.3 The runoff model

The HBV (Hydrologiska Byråns Vattenavdelning) model (Bergström, 1976; Lindström et al., 1997) is a bucket-type hydrologic model for simulating continuous runoff series. Model inputs are daily rainfall and air temperature, as well as daily potential evaporation values. Hydrologic processes are represented by four different routines corresponding to snow, soil water, groundwater, and runoff routing, with a combined total of 16 parameters. In the snow routine, snow accumulation and snowmelt are calculated by a degree-day method. Snowmelt together with rainfall and potential evaporation are input to the soil-water routine, where the actual evaporation and the groundwater recharge are computed based on the soil-moisture storage. The groundwater (or response) routine consists of a connected shallow and deep groundwater reservoir and simulates peak flow, intermediate runoff, and baseflow. These three runoff components are taken together and transformed by a triangular weighting function during the routing process to calculate the runoff at the catchment outlet. Runoff can be modeled in a semi-distributed way by separating a catchment into elevation bands. Thereby, the snow and soil-water routines are calculated for each elevation band, whereas the groundwater storage and the runoff routing routines are treated as a lumped representation of the entire catchment. HBV exists in different versions, whereby the general structure of the model remains the same. The version applied in this study is HBV-light (Seibert and Vis, 2012). Like for all bucket-type models, parameters in the HBV model cannot be determined a priori: they are identified by model calibration instead. More detailed information on the HBV model can be found in Bergström (1976), Lindström et al. (1997), and Seibert and Vis (2012).

2.4 Modeling approach

2.4.1 Model setup

For each of the 25 catchments the number of elevation bands was defined by splitting the catchment into elevation zones of 200 m. Elevation zones covering less than 5 % of the catchment area were merged with the adjacent elevation zone. For the resulting elevation bands, air temperature and rainfall were computed with a lapse rate of 0.6 °C per 100 m and 10 % per 100 m, respectively. Potential evaporation was assumed to be uniform over the whole catchment.

Model simulations were run for two time periods, one lasting from the hydrological years (1 October until 30 September) 1984 to 1996 and the other lasting from 1997 to 2009. The approximately 3 years preceding each simulation period (January 1982 to September 1984 and January 1995 to September 1997, respectively) served to establish state variables of the model. A warm-up period was needed to ensure that the different state variables at the beginning of the simulation period were consistent with the preceding meteorological conditions and parameter values. The two simulation periods were used for model calibration and validation. For calibration, a genetic algorithm (Seibert, 2000) was used and the range of possible parameter values was specified based on previous studies (Lindström et al., 1997; Seibert, 1999; Table 2 in Vis et al., 2015). The 100 independent calibration trials allowed us to account for parameter uncertainty or equifinality (Beven and Freer, 2001) and resulted in a set of 100 calibrated parameter sets for each objective function (Fig. 2).

Table 1. Description of streamflow characteristics used to calibrate the runoff model (adapted from Knight et al., 2014; U.S. Geological Survey, 2014) ($mm\,d^{-1}$: millimeters per day; –: no units; a^{-1}: per annum; %: percent).

Streamflow characteristic	Abbreviation	Further explanation	Flow condition	Unit
		Magnitude		
Mean annual runoff	MA41	Mean annual daily runoff	Mean flow	($mm\,d^{-1}$)
Maximum October runoff	MH10	Mean of October runoff maxima for each year	High flow	($mm\,d^{-1}$)
Lowest 15 % of daily runoff	E85	Daily mean runoff that is exceeded 85 % of the time for the period of record	Low flow	($mm\,d^{-1}$)
Rate of runoff recession	RA7	Median change in log of runoff for days in which the change is negative across the period of record	Mean flow	($mm\,d^{-1}$)
		Ratio		
Average 30-day maximum runoff	DH13	Mean annual maximum of a 30-day moving average runoff divided by the median for the entire record	High flow	(–)
Baseflow	ML20	Ratio of total baseflow to total flow. Baseflow is the minimum flow magnitude in a 5-day window if 90 % of that minimum flow magnitude is less than the minimum flow magnitude of the 5 day window before and after the considered window	Low flow	(–)
Stability of runoff	TA1	Measure of the constancy of a flow regime by dividing daily flows into predetermined flow classes. The 11 flow classes capture flow ranging from flow less than 0.1 times the logarithmic mean flow to flow more than 2.25 times the logarithmic mean flow.	Mean flow	(–)
		Frequency		
Frequency of moderate floods	FH6	Average number of high-flow events per year that are equal to or greater than 3 times the median annual flow for the period of record	High flow	(a^{-1})
Frequency of larger floods	FH7	Average number of high-flow events per year that are equal to or greater than 7 times the median annual flow for the period of record	High flow	(a^{-1})
		Variability		
Variability of March runoff	MA26	Standard deviation for March runoff over the period of record divided by the mean runoff for March over the period of record	Mean flow	(%)
Variability in high-flow pulse duration	DH16	Standard deviation for the yearly average high-flow pulse duration (daily flow greater than the 75th percentile) divided by the mean of the yearly average high-flow pulse duration multiplied by 100	High flow	(%)
Variability of low-flow pulse count	FL2	Standard deviation for the average number of yearly low-flow pulses (daily flow less than the 25th percentile) divided by the mean low-flow pulse counts multiplied by 100	Low flow	(%)
		Date		
Timing of annual minimum runoff	TL1	Julian date of annual minimum flow occurrence	Low flow	(Julian day)

2.4.2 Choice of objective functions for model calibration

The complete model calibration process was conducted for 25 catchments and using data from all five different types of objective functions (see Table 2 for the exact equations) that focused on different aspects of the hydrograph. In the first step, model parameters were constrained by maximizing the model efficiency (R_{eff}, Nash and Sutcliffe, 1970). The model efficiency is the most widely used objective function

Table 2. Objective functions used in model calibration. Objective functions were calculated with observed (obs) and simulated (sim) runoff (Q) or SFCs (I).

Objective function	Abbreviation	Definition	Optimal value
Model efficiency	R_{eff}	$1 - \frac{\sum (Q_{\text{obs}} - Q_{\text{sim}})^2}{\sum (Q_{\text{obs}} - \overline{Q_{\text{obs}}})^2}$	1
Efficiency for each individual SFC[1]	I_{Single}	$1 - \frac{\lvert I_{\text{obs}} - I_{\text{sim}} \rvert}{I_{\text{obs}}}$	1
SFC and model efficiency	$I_{\text{Single\_Reff}}$	$0.5\left(I_{\text{Single}} + R_{\text{eff}}\right)$	1
Efficiency for the selected SFCs[2]	I_{Multi}	$\frac{1}{n}\left(I_{\text{Single}_1} + \ldots + I_{\text{Single}\_n}\right)$	1
SFCs and model efficiency	$I_{\text{Multi\_Reff}}$	$\frac{n-1}{n} I_{\text{Multi}} + \frac{1}{n} R_{\text{eff}}$	1

[1] For each of the 13 SFCs a specific I_{Single} exists.
[2] I_{Multi} consists of the n most robust and informative SFCs.

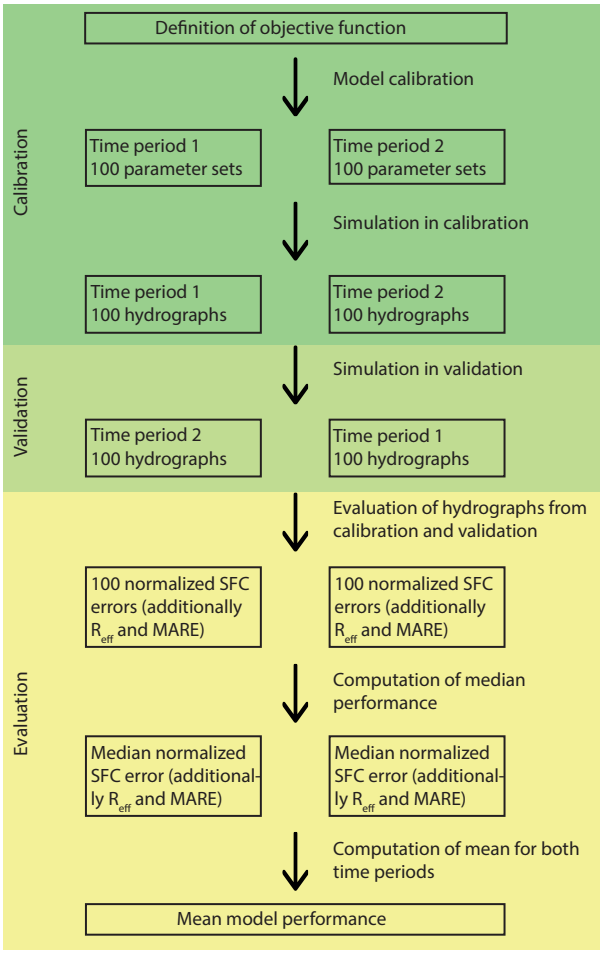


Figure 2. Flow chart of the modeling approach consisting of calibration, validation, and evaluation in time period 1 (1984–1996) and time period 2 (1997–2009) and completed for each of the five objective function types R_{eff}, I_{Sinlge}, $I_{\text{Single\_Reff}}$, I_{Multi}, and $I_{\text{Multi\_Reff}}$.

in hydrological modeling, and it served as a benchmark for the objective functions that included SFCs. Model calibration with R_{eff} tends to reduce simulation errors in magnitude and timing of high-flow conditions at the expense of errors in low-flow conditions (Legates and McCabe, 1999; Krause et al., 2005).

Next, a new efficiency measure that consisted of one single SFC (I_{Single}) was defined to explicitly incorporate individual SFCs into model calibration (Table 2). Each of the 13 selected SFCs was used separately for model calibration, resulting in 13 versions of I_{Single}. Additionally, each SFC efficiency measure was combined with R_{eff}, whereby both metrics were equally weighted ($I_{\text{Single\_Reff}}$). The use of a single SFC as the objective function allowed calibration to focus on a specific aspect of the hydrograph, while adding R_{eff} helped to improve the overall shape of the hydrograph, including the magnitude and timing of events.

Based on the results from the individual SFCs, an objective function consisting of equally weighted normalized SFCs was defined (I_{Multi}, Table 2). This "SFC-based" efficiency measure was again combined with R_{eff} ($I_{\text{Multi\_Reff}}$). For the resulting combined objective function, the same weights were assigned to each metric to make sure the individual SFCs had sufficient influence on the model calibration and were not dominated by R_{eff}. The number of SFCs constituting I_{Multi} was not previously fixed. Instead, a minimum number of SFCs was selected so that the resulting objective function was both robust and informative. These two requirements for the objective function could be achieved by only including SFCs that are robust and informative. A SFC was considered robust when the SFC calculated from a model simulation with I_{Single} had relatively small errors over the full range of catchments in both validation time periods compared to other SFCs. A SFC was regarded as being informative when it also yielded relatively good simulations for other SFCs. The robustness and information value of a SFC were therefore assessed relative to other SFCs, enabling acceptable trade-off solutions for all SFCs, with a minimum

Table 3. Performance measures used in model evaluation. Performance measures were calculated with observed (obs) and simulated (sim) runoff (Q) or SFCs (I).

Performance measure	Abbreviation	Definition	Optimal value
Model efficiency	R_{eff}	$1 - \frac{\sum (Q_{\text{obs}} - Q_{\text{sim}})^2}{\sum (Q_{\text{obs}} - \overline{Q_{\text{obs}}})^2}$	1
Mean absolute relative error[1]	MARE	$1 - \frac{1}{n} \sum \frac{\lvert Q_{\text{obs}} - Q_{\text{sim}} \rvert}{Q_{\text{obs}}}$	1
Normalized SFC error[2]	nSFC	$\frac{I_{\text{obs}} - I_{\text{sim}}}{R_{\text{obs}}}$	0

[1] n is the number of days.
[2] R is the range of possible values of a SFC for the respective catchment.

number of SFCs being potentially representative for (most of) the 13 SFCs.

2.4.3 Evaluation of model performance

Model performance in calibration and validation was evaluated by means of normalized SFC error, R_{eff}, and mean absolute relative error (MARE) (see Table 3 for the exact equations). These evaluation criteria were calculated for all 100 runoff simulations based on the five different types of objective functions in both validation time periods and for all 25 catchments. For the interpretation of the results, the median model efficiency of each objective function, validation period, and catchment was selected as the representative value for the model efficiency distribution. Simulation uncertainty stemming from the 100 parameter sets was assessed by a two-sided binomial test with the null hypothesis that the probability for overestimation and underestimation of a SFC is equal to 50 %.

As there are significant differences in the SFC ranges, a normalization was needed that allowed comparison of the different SFCs. Instead of normalizing in terms of relative error, an approach was applied that normalizes the SFC estimation error. The normalization of a SFC was computed as the absolute simulation error divided by the range of possible values for that SFC in the respective catchment (Table 3). To calculate these SFC ranges, 10 000 Monte Carlo simulations were run for each respective catchment using randomly chosen parameter values from the previously identified parameter space (Lindström et al., 1997; Seibert, 1999; Table 2 in Vis et al., 2015). The Monte Carlo simulations represented the potential variation in a certain SFC if no information was available to constrain the runoff model. The range was then calculated as the difference between the 10th and 90th percentiles of the simulated SFC values.

3 Results

The HBV model was capable of reproducing the observed runoff for the study catchments reasonably well. Model calibration on R_{eff} resulted in R_{eff} values between 0.68 and 0.89 with a median of 0.79. The corresponding R_{eff} values in validation ranged from 0.62 to 0.86 with a median of 0.77.

3.1 The use of single SFCs as objective functions in model calibration

3.1.1 Estimation accuracy using SFC-specific model calibrations

Model calibration results for the 13 SFCs confirmed that HBV-light is capable of estimating different SFCs with a high level of precision if the respective SFC was used as an objective function (I_{Single}) for model calibration (the 13 absolute nSFCs varied between 0.000 and 0.005 for calibrations with I_{Single}). Both I_{Single} and the combined objective function I_{Single_Reff} clearly outperformed model calibrations based on R_{eff} with regard to the estimation of SFCs (Fig. 3a). However, calibration with I_{Single} yielded poor model performances when evaluated in terms of R_{eff}, whereas R_{eff} efficiencies of calibrations with either $I_{\text{Single\_Reff}}$ or R_{eff} were comparable (Fig. 3a).

Validation results (Fig. 3b) exhibited a similar pattern in model performance to the calibration results. The median absolute normalized error of the 13 SFCs was relatively low for model runs based on the objective functions I_{Single} and $I_{\text{Single\_Reff}}$ compared to model calibration with R_{eff}. The comparable SFC estimation accuracy of I_{Single} and $I_{\text{Single\_Reff}}$ that often outperformed model simulations with R_{eff} confirms the value of SFCs for model calibration aiming at a respective SFC. An exceptional behavior can be observed for MH10, where the estimation accuracy was negatively affected by a calibration based on the SFC itself (Fig. 5a–c).

3.1.2 How informative is a SFC for estimating any SFC?

The calibrations for all 13 versions of I_{Single} and $I_{\text{Single\_Reff}}$ resulted in a total in 26 different runoff simulations that were evaluated by calculating the normalized SFC error for the calibration and validation periods. The SFC TA1 (stability of runoff; Fig. 4a and b) was selected as a representative example to illustrate that the use of SFCs as a single objective

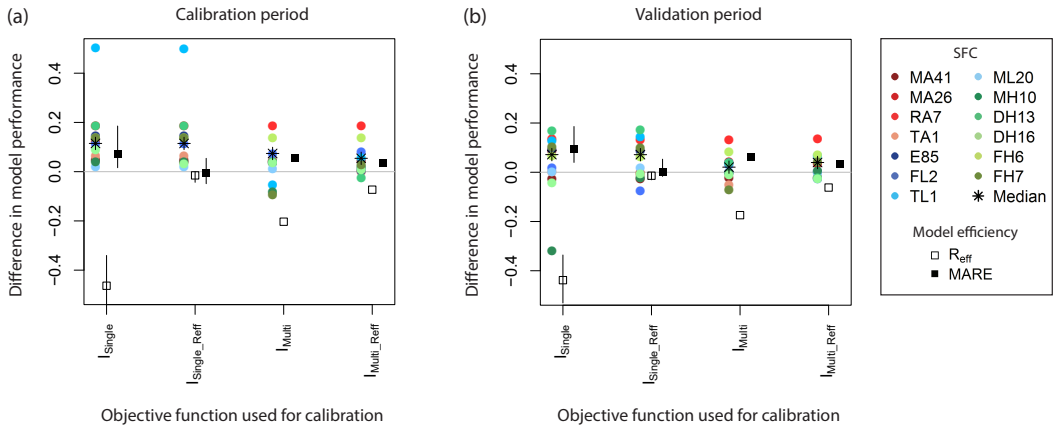


Figure 3. Model performance in **(a)** calibration and **(b)** validation in terms of absolute normalized SFC errors (nSFC), R_{eff}, and MARE depending on the objective function used in calibration. Model performance is shown as the difference between a model calibration with R_{eff} and model calibrations with I_{Single}, $I_{\text{Single\_Reff}}$, I_{Multi}, or $I_{\text{Multi\_Reff}}$ (positive values indicate that model calibration with I_{Single}, $I_{\text{Single\_Reff}}$, I_{Multi}, or $I_{\text{Multi\_Reff}}$ resulted in better model performance than model calibration with R_{eff}; negative values indicate that model calibration with I_{Single}, $I_{\text{Single\_Reff}}$, I_{Multi}, or $I_{\text{Multi\_Reff}}$ resulted in poorer model performance than model calibration with R_{eff}). Model performance values correspond to the median of the 25 catchments and the mean of both modeling time periods.

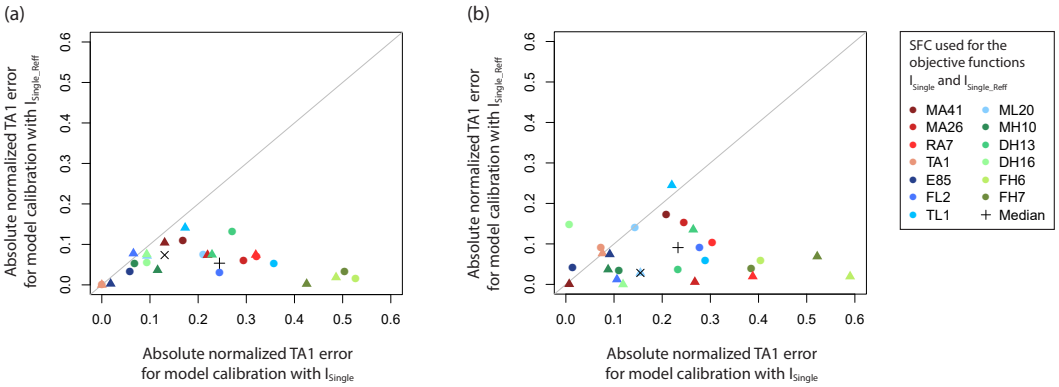


Figure 4. Absolute normalized TA1 error (nSFC) in **(a)** calibration and **(b)** validation calculated from model calibrations with the objective functions I_{Single} and $I_{\text{Single\_Reff}}$. Absolute normalized SFC errors correspond to the median of the 25 catchments and are shown separately for both modeling time periods (triangles for period 1, 1984–1996, and circles for period 2, 1997–2009). The x and plus symbols represent the median of periods 1 and 2, respectively. (Absolute normalized TA1 error for model calibrations with the objective function R_{eff} was 0.08, period 1, and 0.05, period 2, in calibration and 0.002, period 1, and 0.15, period 2, in validation.)

function (I_{Single}) generally resulted in poor SFC estimates for those SFCs not included in I_{Single} in both model calibration and validation when compared to model calibrations with $I_{\text{Single\_Reff}}$ or R_{eff}. Estimation accuracies from calibrations with $I_{\text{Single\_Reff}}$ and R_{eff} were often of comparable magnitude. Error magnitudes from the three described objective function types (I_{Single}, $I_{\text{Single\_Reff}}$, and R_{eff}) could vary considerably between time periods (illustrated by triangles and circles, respectively, in Fig. 4a and b).

3.2 The use of multiple SFCs for model calibration

Figure 6a shows simulation results for the objective function I_{Single} for all 25 catchments and both modeling time periods. The five SFCs with the highest robustness (less variability in error; Fig. 6a) were RA7, ML20, FH6, E85, and MA41. All five of these SFCs could be used for the objective function I_{Multi}; however, E85 (lowest 15 % of daily runoff) was discarded as potential SFC for I_{Multi} because of its redundant information with ML20 (baseflow). The information value of the remaining 4 SFCs for each of the 13 SFCs is presented in Fig. 6b. All 13 SFCs were relatively well simulated by model calibrations with I_{Single} of either RA7, ML20, FH6, or MA41 (colored circles in Fig. 6b) compared to calibrations with other SFCs.

Median estimates of the 13 SFCs in the calibration period were slightly lower when the model was calibrated with

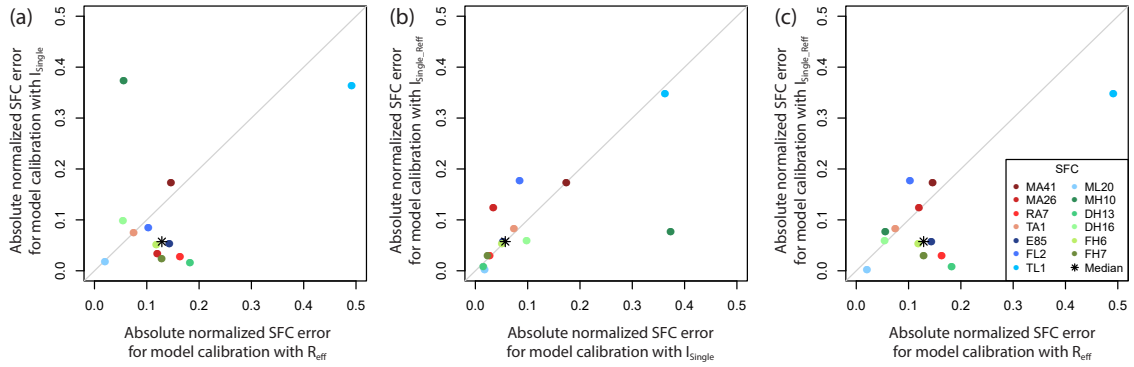


Figure 5. Comparison of absolute normalized SFC errors (nSFC) in validation calculated from model calibrations with the objective functions R_{eff}, I_{Single}, and I_{Single_Reff}. Absolute normalized SFC errors correspond to the median of the 25 catchments and the mean of both modeling time periods.

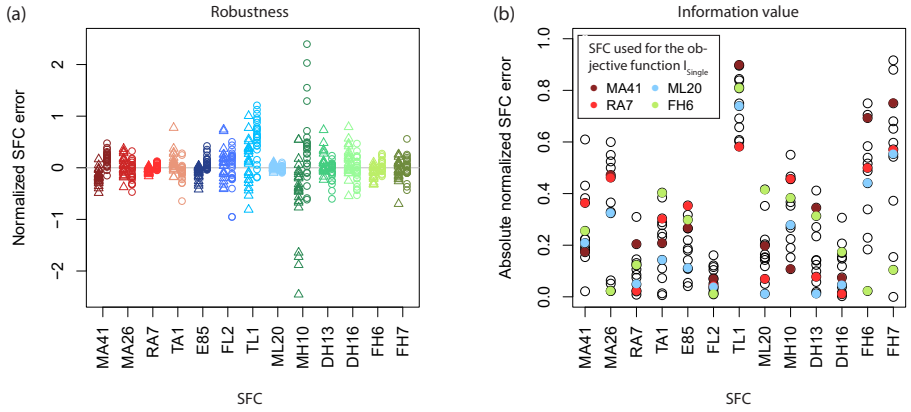


Figure 6. (a) Robustness: normalized SFC errors (nSFC) in validation calculated from model calibrations with the objective function I_{Single} for the respective SFC. Values are shown for all 25 catchments and both modeling time periods (triangles for period 1, 1984–1996, and circles for period 2, 1997–2009). **(b)** Information value: absolute normalized SFC errors (nSFC) in validation calculated from model calibrations with all 13 objective functions I_{Single}. Model performance values correspond to the median of the 25 catchments and the mean of both modeling time periods. Each open circle represents 1 of the 13 SFCs used for I_{Single}. The colored circles refer to the final selection of SFCs for the objective function I_{Multi}.

I_{Multi} rather than I_{Multi_Reff}. Both of these objective functions led to better model performance for SFCs than calibrating with R_{eff} alone (Fig. 3a). Model performance for the validation period with I_{Multi_Reff} had a lower median error for SFCs than the error associated with using I_{Multi} as an objective function (Fig. 3b). The comparison of I_{Multi} and I_{Multi_Reff} for all SFCs separately (Fig. 7a) revealed that for most SFCs both objective functions resulted in similar estimates. I_{Single_Reff} was better for estimating SFCs than I_{Multi_Reff}, especially for SFCs not included in the I_{Multi_Reff} objective function (Fig. 7b). Comparing simulations from I_{Multi_Reff} and R_{eff} revealed a smaller median error of the SFCs when calibrating with I_{Multi_Reff} (Figs. 3b and 7c). Yet, for most SFCs not explicitly incorporated into the objective function I_{Multi_Reff}, the objective function R_{eff} performed equally well or slightly better than I_{Multi_Reff} (Fig. 7c).

3.3 Estimation accuracy for SFCs

Figure 8 provides an overview (median of all 25 catchments) of how well SFCs were simulated by presenting the results for both modeling time periods and all five objective function types. Error magnitudes ranged between −25 and 25 % for the majority of SFCs. Considerably higher estimation accuracy was achieved for ML20 (−5 to 2 %), whereas estimation accuracies were lowest for MH10 and TL1, with error magnitudes up to 40 and 77 %, respectively. For some SFCs (e.g., MA26 and TL1) the error tended to be higher in one of the two modeling time periods, whereas for other SFCs (e.g., RA7 and MH10) the objective function had a distinct influence on the error magnitude. There was no evidence that the estimation accuracy depends on flow components (magnitude, ratio, frequency, variability, and date) or flow conditions (low, medium, and high flow).

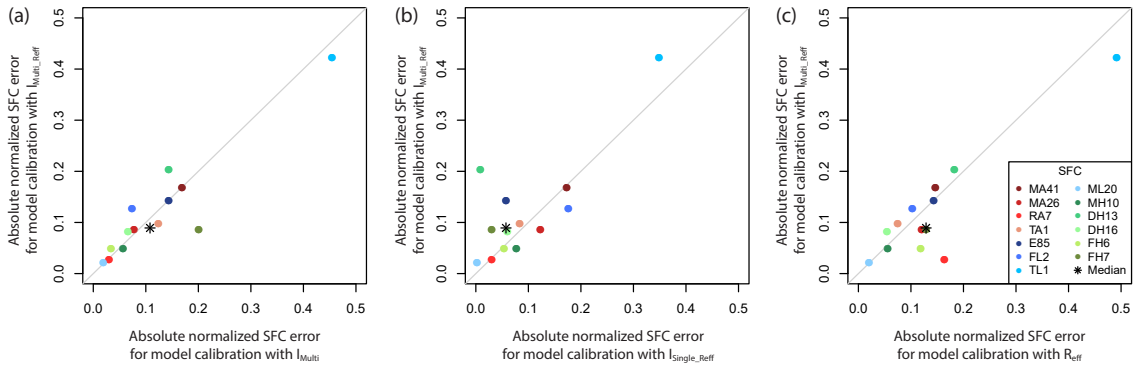


Figure 7. Comparison of absolute normalized SFC errors (nSFC) in validation calculated from model calibrations with the objective functions R_{eff}, I_{Single_Reff}, I_{Multi}, and I_{Multi_Reff}. Absolute normalized SFC errors correspond to the median of the 25 catchments and the mean of both modeling time periods.

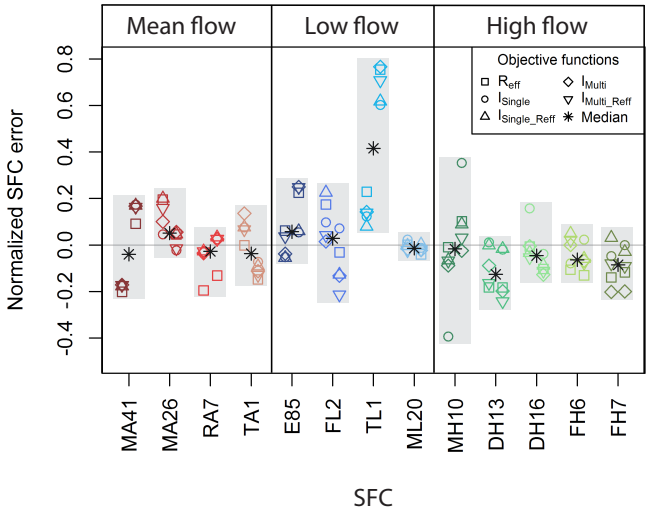


Figure 8. Normalized SFC errors (nSFC) in validation depending on the objective function used in calibration. Model performance values correspond to the median of the 25 catchments and are shown for both modeling time periods (period 1, 1984–1996, on the left side and period 2, 1997–2009, on the right side).

The median error (illustrated by stars in Fig. 8) was used for the evaluation of the underestimation or overestimation of SFCs. Among the tested SFCs, an underestimation was observed for all five SFCs representing high-flow conditions as well as for three of four mean-flow-related SFCs. With one exception, low-flow SFCs were overestimated. This overall pattern was less evident when evaluating each objective function and time period separately (Figs. 8 and 9). The SFCs DH16 and MH10 indicate two typically observed deviations in the overall pattern. DH16 is an example of a SFC that could be regarded as being clearly underestimated by the model, because of its negative bias in 9 out of 10 cases (median values in Fig. 9a). However, for objective functions or modeling time periods with a low magnitude in the median bias, the underestimation of the SFC was not statistically significant. Even in the case of a median pointing to statistically significant underestimation, there might be a substantial number of catchments for which DH16 was overestimated. A second commonly observed phenomenon is shown by the SFC MH10 (Fig. 9b). While MH10 had mostly small but statistically significant median errors, there were many catchments with considerably higher errors. Although MH10 was the most extreme example, it illustrates that small median errors do not guarantee good results for all catchments.

4 Discussion

4.1 On the importance of the choice of the objective function

The results demonstrated that the objective function used for model calibration strongly influences the estimation accuracy of SFCs. This finding confirms the findings of previous studies (e.g., Hingray et al., 2010; Westerberg et al., 2011; Murphy et al., 2013; Olsen et al., 2013; Pfannerstill et al., 2014; Shrestha et al., 2014; Caldwell et al., 2015; Vis et al., 2015) and points out the importance of making a careful choice of the objective function for model calibration. The benefit of optimizing one specific SFC lies in the relatively accurate estimation of the respective SFC compared to a calibration with R_{eff} or a multi-SFC objective function. Model calibration on one single SFC clearly emphasizes the hydrograph aspects of the selected SFC possibly neglecting an adequate representation of other hydrograph characteristics. This implies that calibrations with I_{Single} can lead to poor model performance for SFCs not included in the objective function. The fact that a calibration with R_{eff} and a calibration with multiple SFCs lead to comparable estimates for most SFCs indicates that the main hydrological processes of the catchments are similarly well represented with the two approaches. Considering that SFCs not incorporated into the objective function

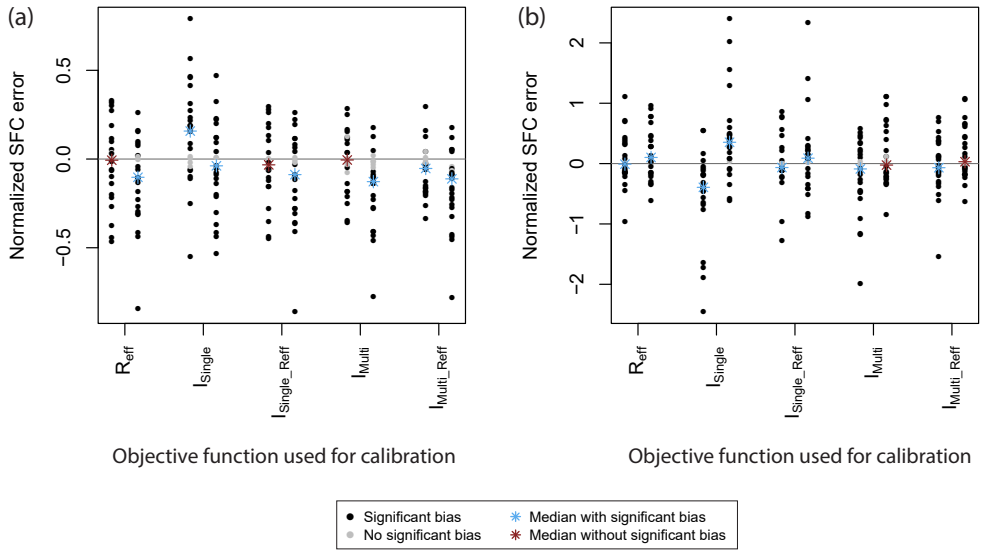


Figure 9. (a) Normalized DH16 errors (nSFC) and **(b)** normalized MH10 errors (nSFC) in validation depending on the objective function used in calibration. Normalized SFC errors are shown for all 25 catchments and for both modeling time periods (period 1, 1984–1996, on the left side and period 2, 1997–2009, on the right side). Colors indicate the significance of the results assessed by a two-sided binomial test at a confidence level of 0.95. Note the difference in the y axis.

I_{Multi} showed little change compared to calibrations with R_{eff} brings into question the benefit of including SFCs in model calibration instead of applying a traditional calibration approach when aiming at estimating a suite of SFCs. This is surprising because the SFCs selected for I_{Multi} or $I_{\text{Multi\_Reff}}$ provide information on high flows, recession rate, percentage of baseflow, and annual runoff volume, and therefore should help in constraining the model with respect to different important runoff processes. These results are different from those of Yilmaz et al. (2008) and Pfannerstill et al. (2014), whose multi-metric runoff model calibration resulted in an improved general shape of the hydrograph. Although their calibration approach was mainly based on various segments of the flow duration curve, it is unclear why the conclusions differ that much. From the above discussion it becomes evident that calibrating a runoff model for estimating many different SFCs from one single hydrograph is a trade-off between finding a parameterization that is general enough to represent different aspects of the hydrograph and that simultaneously emphasizes specific SFCs. These trade-off situations are common as perfect model parameterizations are usually not possible due to a variety of uncertainty sources, such as model structural uncertainty and input and runoff data uncertainty (Beven, 2016).

A noticeable result from the current study is the distinct difference in model performance in calibration and validation when using the objective function I_{Single}. While almost perfect fits are achieved in calibration for all catchments and SFCs, model errors tend to be much higher in validation, with a considerable spread between catchments as well as a clear difference depending on the SFC. This observation confirms that the model is able to simulate the SFCs well, but also outlines that a good model calibration does not imply robust simulations in validation. In general, it seems that SFCs that are strongly related to physical catchment properties (e.g., rate of streamflow recession) are the most robust, followed by SFCs representing an average flow condition with a moderate robustness. SFCs that are a measure of more extreme high-flow conditions are the least robust, possibly because these conditions are subject to inter-annual weather changes and are more difficult to model due to their dynamic behavior. A low robustness could also indicate that the model structure might be suboptimal for some catchments.

The two least robust SFCs are MH10 and TL1. MH10 simulations with I_{Single} yield by far the poorest results of all objective function types, with very large normalized error in both positive and negative directions. In comparison, the high estimation errors for TL1 depend on the modeling time period. The high estimation errors for TL1 in period 2 stem from years where the minimum runoff was simulated in late winter while the observed minimum was in late fall. By visually analyzing the temperature and runoff time series, it can be hypothesized that such model simulations mainly happened in years with successive weeks of continuously little precipitation during late winter. Such prolonged drier periods occurred more often in one of the two modeling time periods and thus evoked the distinct bias in model accuracy depending on the simulation period. Both TL1 and MH10 are calculated from a single value per year, as opposed to, e.g., RA7, which is based on all recessions. In model calibration, many parameter sets are derived that perfectly simulate this single value. However, a good simulation of either TL1 or

MH10 is not so much dependent on an accurate representation of dominant runoff processes. Thus, model results for the validation period using input data of identical quality can fail to accurately simulate either SFC because of parameter sets "tuned" to the data as opposed to being based on modeling the process.

4.2 Model performance regarding SFCs

The runoff model tends to underestimate SFCs related to mean and high-flow conditions, while SFCs representing low-flow conditions are generally overestimated. These results are consistent with those of Olsen et al. (2013), Caldwell et al. (2015), Vis et al. (2015), and Kiesel et al. (2017) and can partly be explained by the model behavior characterized by a less pronounced runoff response to precipitation events but increased groundwater discharge to the stream during drier periods compared to the observed data (Vis et al., 2015). The observations that average flow conditions are better simulated than extremes (Caldwell et al., 2015; Vis et al., 2015) or that high-flow-related SFCs are more accurately estimated than those related to low flow (Shrestha et al., 2014; Ryo et al., 2015) cannot be confirmed with our results. None of these earlier studies explicitly included SFCs in model calibration and the deviating results could be attributed to the differing approaches to defining the objective function(s). This presumption is supported by the previously described differences in results of Vis et al. (2015), although they applied the same runoff model, catchments, and SFCs.

4.3 How to select SFCs for a multi-index calibration approach

The current study supports the assumption that including SFCs in model calibration helps to preserve most hydrograph aspects relevant to those SFCs. Thus, an objective function based on several SFCs is expected to result in a hydrograph from which a suite of SFCs can be calculated. Not knowing which SFCs will be relevant for a given study, a guideline as to which SFCs the model calibration could be based on would be helpful. The first step towards a guideline consists of selecting SFCs that are potentially valuable for model calibration. This selection was based on the concept of robustness and information value of SFCs, which is comparable to the approach used by Euser et al. (2013), who assessed the realism of model structures. Like Euser et al. (2013), results from the current study indicated that high robustness was not necessarily related to high information value, emphasizing the importance of selecting SFCs by jointly evaluating robustness and information value. The concept of information value and robustness favors simulations that preserve important hydrograph characteristics, as can be seen from the slightly improved median estimation accuracy of SFCs with the objective functions I_{Multi} or $I_{\text{Multi\_Reff}}$ compared to estimations with R_{eff} only.

A model calibrated on certain flow conditions (low, medium, and high flow) is beneficial for SFCs representing these flow conditions (see, e.g., Murphy et al., 2013), so it was hypothesized that the information value of the selected SFCs is highest for SFCs belonging to the same group of flow conditions. The confirmation of this hypothesis would allow us to draw general conclusions about a minimum number of SFCs required for model calibration. Surprisingly the results did not reveal any pattern related to flow conditions and thus no recommendation for the final selection of SFCs can be made. It seems that the selection of SFCs for an informative and robust objective function depends on the type and the combination of SFCs one is interested in. Since this study was based on a limited number of SFCs it could be interesting to test the hypothesis by analyzing a greater number of SFCs. Testing a larger number of SFCs might reveal relations that are difficult to see with a small sample. Furthermore, more knowledge about the effect of single SFCs or the combination of SFCs used as objective functions on runoff simulations could be gained by using synthetic data and a modeling approach where an excellent hydrograph fit is possible (e.g., "HBV-land" in Seibert and Vis, 2012).

4.4 Objective functions, their estimation accuracy, and consequences for practical applications

The emphasis of SFC-related modeling studies changed from estimating single SFCs to simulating a suite of SFCs (Olden and Poff, 2003). The modeling design of this study combined both approaches for the same SFCs and catchments and thus enabled a direct comparison of the results. Ideally, the runoff model could be calibrated to simulate a hydrograph for each catchment from which any SFC can be calculated. Such an approach ensures a relatively small calibration effort, which is especially valuable if one is interested in modeling many catchments and/or various scenarios. However, results indicate that SFCs related to a more generally calibrated model (e.g., R_{eff}, I_{Multi}, or $I_{\text{Multi\_Reff}}$) are less accurate than when they are estimated from hydrographs based on targeted model calibrations (e.g., I_{Single} or $I_{\text{Single\_Reff}}$). This fact has substantial implications for the later application of simulated SFCs in decision-support systems for integrated resource management. As stated by Carlisle et al. (2010), with high errors in SFC estimates, only considerable flow departures from natural conditions can be detected. Also, inaccurate SFC values can impede the generation of more robust flow alteration–ecosystem change relationships that are ultimately needed for sustainable flow management guidelines (Arthington et al., 2006; Poff and Zimmermann, 2010; Gillespie et al., 2015; Cartwright et al., 2017).

As with regional statistical approaches, incorporating SFCs into model objective functions implies that a modeler knows which SFCs are relevant and that the model must be recalibrated if one is interested in additional SFCs. The advantage of runoff models over multivariate regressions and

observed streamflow series includes their use for climate scenario analysis or for simulating runoff in ungauged catchments, with the latter being one of the ultimate aims in the ELOHA framework (Poff et al., 2010). Modeling SFCs gets even more challenging when moving from a gauged to an ungauged catchment. An appropriate calibration strategy targeted to the main simulation goal is crucial for any subsequent regionalization.

4.5 Choice of the runoff model for estimating SFCs

When comparing SFCs estimated from simulations of different runoff models, the question can be raised whether the results depend on the selected model. This question is especially important for resource managers who need to make decisions based on model results from different studies (Caldwell et al., 2015). A comparison of runoff models with different spatial scales that rely on different data inputs was conducted by Caldwell et al. (2015). Their results do not indicate that a certain runoff model is more suited for predicting SFCs than others, but rather that the calibration process probably has as much influence as the model structure. Thus, it can be assumed that the conclusions of this study would be similar if a different calibrated runoff model was applied.

5 Conclusions

In this study, we evaluated the value of using SFCs for the calibration of a runoff model used to estimate SFCs. The results suggest that the choice of the objective function used for model calibration strongly influences the estimation accuracy of SFCs. While the model was capable of correctly simulating any of the tested SFCs, a good reproduction of a particular SFC was generally achieved when this SFC was included in the objective function. SFC estimates from model simulations with an objective function consisting of a representative selection of SFCs resulted in comparable accuracies to the estimates from model runs based on the commonly used model efficiency when evaluated against SFCs not included in the objective function. Estimates of SFCs that are less dependent on the short-term weather input or SFCs representing average flow conditions were more robust than other SFCs. Since the results imply that one has to consider significant uncertainties when simulated time series are used to derive SFCs that were not included in the calibration, we strongly recommend calibrating the runoff model explicitly for the SFCs of interest.

Data availability. Data used in this study are available at the U.S. Department of Commerce (2007a, b) and the U.S. Geological Survey (2016a, b).

Author contributions. SP, MV, RK, and JS designed this study based on a previous collaboration; MV performed the runoff simulations; SP analyzed the results that were discussed with all co-authors. Writing of the paper was led by SP with contribution of all co-authors.

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. This paper is a product of discussions and activities that took place at the U.S. Geological Survey John Wesley Powell Center for Analysis and Synthesis as part of the workgroup focusing on Water Availability for Ungauged Rivers (https://powellcenter.usgs.gov/). Funding for this research was provided by the U.S. Geological Survey Cooperative Water Program and the University of Zurich. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. We thank the reviewers Björn Guse, Oddbjørn Bruland, and Sjur Kolberg for their constructive and detailed comments that helped to improve the quality of our manuscript.

Edited by: Dimitri Solomatine
Reviewed by: Oddbjørn Bruland, Björn Guse, and Sjur Kolberg

References

Abell, R. A., Olson, D. M., Dinerstein, E., Hurley, P. T., Diggs, J. T., Eichbaum, W., Walters, S., Wettengel, W., Allnutt, T., Loucks, C. J., and Hedao, P. (Eds.): Freshwater ecoregions of North America: A conservation assessment, Island Press, Washington, DC, USA, 2000.

Arthington, A. H., Bunn, S. E., Poff, N. L., and Naiman, R. J.: The challenge of providing environmental flow rules to sustain river ecosystems, Ecol. Appl., 16, 1311–1318, https://doi.org/10.1890/1051-0761(2006)016[1311:TCOPEF]2.0.CO;2, 2006.

Bergström, S.: Development and application of a conceptual runoff model for Scandinavian catchments, SMHI, Norrköping, Sweden, No. RHO 7, 134 pp., 1976.

Beven, K.: Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication, Hydrolog. Sci. J., 61, 1652–1665, https://doi.org/10.1080/02626667.2015.1031761, 2016.

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol. 249, 11–29, https://doi.org/10.1016/S0022-1694(01)00421-8, 2001.

Caldwell, P. V., Kennen, J. G., Sun, G., Kiang, J. E., Butcher, J. B., Eddy, M. C., Hay, L. E., LaFontaine, J. H., Hain, E. F., Nelson, S. A. C., and McNulty, S. G.: A comparison of hydrologic models for ecological flows and water availability, Ecohydrology, 8, 1525–1546, https://doi.org/10.1002/eco.1602, 2015.

Carlisle, D. M., Falcone, J., Wolock, D. M., Meador, M. R., and Norris, R. H.: Predicting the natural flow regime: models for as-

sessing hydrological alteration in streams, River Res. Appl., 26, 118–136, https://doi.org/10.1002/rra.1247, 2010.

Cartwright, J., Caldwell, C., Nebiker, S., and Knight, R.: Putting flow–ecology relationships into practice: A decision-support system to assess fish community response to water-management scenarios, Water, 9, 196, https://doi.org/10.3390/w9030196, 2017.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H. H. G.: A framework to assess the realism of model structures using hydrological signatures, Hydrol. Earth Syst. Sci., 17, 1893–1912, https://doi.org/10.5194/hess-17-1893-2013, 2013.

Gillespie, B. R., Desmet, S., Kay, P., Tillotson, M. R., and Brown, L. E.: A critical analysis of regulated river ecosystem responses to managed environmental flows from reservoirs, Freshwater Biol., 60, 410–425, https://doi.org/10.1111/fwb.12506, 2015.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F.: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling, J. Hydrol., 377, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003, 2009.

Hailegeorgis, T. T. and Alfredsen, K.: Regional statistical and precipitation-runoff modelling for ecological applications: Prediction of hourly streamflow in regulated rivers and ungauged basins, River Res. Appl., 33, 233–248, https://doi.org/10.1002/rra.3006, 2016.

Hingray, B., Schaefli, B., Mezghani, A., and Hamdi, Y.: Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments, Hydrolog. Sci. J., 55, 1002–1016, https://doi.org/10.1080/02626667.2010.505572, 2010.

Hoos, A. B.: Recharge rates and aquifer hydraulic characteristics for selected drainage basins in middle and east Tennessee, U.S. Geological Survey, Nashville, Tennessee, USA, Water Resources Investigations Report 90–4015, 39 pp., 1990.

Jothityangkoon, C., Sivapalan, M., and Farmer, D. L.: Process controls of water balance variability in a large semi-arid catchment: Downward approach to hydrological model development, J. Hydrol., 254, 174–198, https://doi.org/10.1016/S0022-1694(01)00496-6, 2001.

Kiesel, J., Guse, B., Pfannerstill, M., Kakouei, K., Jähnig, S. C., and Fohrer, N.: Improving hydrological model optimization for riverine species, Ecol. Indic., 80, 376–385, https://doi.org/10.1016/j.ecolind.2017.04.032, 2017.

Knight, R. R., Gregory, M. B., and Wales, A. K.: Relating streamflow characteristics to specialized insectivores in the Tennessee River Valley: A regional approach, Ecohydrology, 1, 394–407, https://doi.org/10.1002/eco.32, 2008.

Knight, R. R., Gain, W. S., and Wolfe, W. J.: Modelling ecological flow regime: an example from the Tennessee and Cumberland River basins, Ecohydrology, 5, 613–627, https://doi.org/10.1002/eco.246, 2012.

Knight, R. R., Murphy, J. C., Wolfe, W. J., Saylor, C. F., and Wales, A. K.: Ecological limit functions relating fish community response to hydrologic departures of the ecological flow regime in the Tennessee River basin, United States, Ecohydrology, 7, 1262–1280, https://doi.org/10.1002/eco.1460, 2014.

Krause, P., Boyle, D. P., and Bäse, F.: Comparison of different efficiency criteria for hydrological model assessment, Adv. Geosci., 5, 89–97, https://doi.org/10.5194/adgeo-5-89-2005, 2005.

Law, G. S., Tasker, G. D., and Ladd, D. E.: Streamflow-characteristic estimation methods for unregulated streams of Tennessee, U.S. Geological Survey, Reston, Virginia, USA, Scientific Investigations Report 2009–5159, 212 pp., 2009.

Legates, D. R. and McCabe, G. J.: Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation, Water Resour. Res., 35, 233–241, https://doi.org/10.1029/1998WR900018, 1999.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., and Bergström, S.: Development and test of the distributed HBV-96 hydrological model, J. Hydrol., 201, 272—288, https://doi.org/10.1016/S0022-1694(97)00041-3, 1997.

Murphy, J. C., Knight, R. R., Wolfe, W. J., and Gain, W. S.: Predicting ecological flow regime at ungauged sites: A comparison of methods, River Res. Appl., 29, 660–669, https://doi.org/10.1002/rra.2570, 2013.

Nash, J. E. and Sutcliffe, J. V.: River flow forecasting through conceptual models part I – A discussion of principles, J. Hydrol., 10, 282–290, https://doi.org/10.1016/0022-1694(70)90255-6, 1970.

Olden, J. D. and Poff, N. L.: Redundancy and the choice of hydrologic indices for characterizing streamflow regimes, River Res. Appl., 19, 101–121, https://doi.org/10.1002/rra.700, 2003.

Olsen, M., Troldborg, L., Henriksen, H. J., Conallin, J., Refsgaard, J. C., and Boegh, E.: Evaluation of a typical hydrological model in relation to environmental flows, J. Hydrol., 507, 52–62, https://doi.org/10.1016/j.jhydrol.2013.10.022, 2013.

Omernik, J. M.: Ecoregions of the Conterminous United States, Ann. Assoc. Am. Geogr., 77, 118–125, https://doi.org/10.1111/j.1467-8306.1987.tb00149.x, 1987.

Pfannerstill, M., Guse, B., and Fohrer, N.: Smart low flow signature metrics for an improved overall performance evaluation of hydrological models, J. Hydrol., 510, 447–458, https://doi.org/10.1016/j.jhydrol.2013.12.044, 2014.

Poff, N. L. and Zimmerman, J. K.: Ecological responses to altered flow regimes: A literature review to inform the science and management of environmental flows, Freshwater Biol., 55, 194–205, https://doi.org/10.1111/j.1365-2427.2009.02272.x, 2010.

Poff, N. L., Allan, J. D., Bain, M. B., Karr, J. R., Prestegaard, K. L., Richter, B. D., Sparks, R. E., and Stromberg, J. C.: The natural flow regime, BioScience, 47, 769–784, https://doi.org/10.2307/1313099, 1997.

Poff, N. L., Richter, B. D., Arthington, A. H., Bunn, S. E., Naiman, R. J., Kendy, E., Acreman, M., Apse, C., Bledsoe, B. P., Freeman, M. C., Henriksen, J., Jacobson, R. B., Kennen, J. G., Merritt, D. M., O'Keeffe, Y. H., Olden, J. D., Rogers, K., Tharme, R. E., and Warner, A.: The ecological limits of hydrologic alteration (ELOHA): A new framework for developing regional environmental flow standards, Freshwater Biol., 55, 147–170, https://doi.org/10.1111/j.1365-2427.2009.02204.x, 2010.

Richter, B. D., Baumgartner, J. V., Powell, J., and Braun, D. P.: A method for assessing hydrologic alteration within ecosystems, Conserv. Biol., 10, 1163–1174, https://doi.org/10.1046/j.1523-1739.1996.10041163.x, 1996.

Rotstayn, L. D., Roderick, M. L., and Farquhar, G. D.: A simple pan-evaporation model for analysis of climate simulations: Evaluation over Australia, Geophys. Res. Lett., 33, L7715, https://doi.org/10.1029/2006GL027114, 2006.

Ryo, M., Iwasaki, Y., and Yoshimura, C.: Evaluation of spatial pattern of altered flow regimes on a river network using a distributed hydrological model, PloS ONE, 10, e0133833, https://doi.org/10.1371/journal.pone.0133833, 2015.

Sanborn, S. C. and Bledsoe, B. P.: Predicting streamflow regime metrics for ungauged streams in Colorado, Washington, and Oregon, J. Hydrol., 325, 241–261, https://doi.org/10.1016/j.jhydrol.2005.10.018, 2006.

Sawicz, K., Wagener, T., Sivapalan, M., Troch, P. A., and Carrillo, G.: Catchment classification: empirical analysis of hydrologic similarity based on catchment function in the eastern USA, Hydrol. Earth Syst. Sci., 15, 2895–2911, https://doi.org/10.5194/hess-15-2895-2011, 2011.

Seibert, J.: Regionalization of parameters for a conceptual rainfall-runoff model, Agr. Forest Meteorol., 98–99, 279–293, https://doi.org/10.1016/S0168-1923(99)00105-7, 1999.

Seibert, J.: Multi-criteria calibration of a conceptual runoff model using a genetic algorithm, Hydrol. Earth Syst. Sci., 4, 215–224, https://doi.org/10.5194/hess-4-215-2000, 2000.

Seibert, J. and Vis, M. J. P.: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, Hydrol. Earth Syst. Sci., 16, 3315–3325, https://doi.org/10.5194/hess-16-3315-2012, 2012.

Shrestha, R. R., Peters, D. L., and Schnorbus, M. A.: Evaluating the ability of a hydrologic model to replicate hydroecologically relevant indicators, Hydrol. Process., 28, 4294–4310, https://doi.org/10.1002/hyp.9997, 2014.

Tharme, R. E.: A global perspective on environmental flow assessment: Emerging trends in the development and application of environmental flow methodologies for rivers, River Res. Appl., 19, 397–441, https://doi.org/10.1002/rra.736, 2003.

U.S. Department of Commerce: Divisional normals and standard deviations of temperature, precipitation, and heating and cooling degree days 1971–2000 (and previous normals periods), Section 2 precipitation, United States Department of Commerce, Washington, DC, USA, Climatography of the United States No. 85, 2007a.

U.S. Department of Commerce: Divisional normals and standard deviations of temperature, precipitation, and heating and cooling degree days 1971–2000 (and previous normals periods), Section 1 temperature, United States Department of Commerce: Washington, DC, USA, Climatography of the United States No. 85, 2007b.

U.S. Geological Survey: EflowStats R-package, available at: https://github.com/USGS-R/EflowStats (last access: July 2016), 2014.

U.S. Geological Survey: The National Map, 3D Elevation Program Products and Services Web page, available at: http://nationalmap.gov/3DEP/3dep_prodserv.html (last access: November 2015), 2016a.

U.S. Geological Survey: National Water Information System – Web interface, https://doi.org/10.5066/F7P55KJN, 2016b.

Viglione, A., Parajka, J., Rogger, M., Salinas, J. L., Laaha, G., Sivapalan, M., and Blöschl, G.: Comparative assessment of predictions in ungauged basins – Part 3: Runoff signatures in Austria, Hydrol. Earth Syst. Sci., 17, 2263–2279, https://doi.org/10.5194/hess-17-2263-2013, 2013.

Vis, M., Knight, R., Pool, S., Wolfe, W., and Seibert, J.: Model calibration criteria for estimating ecological flow characteristics, Water, 7, 2358–2381, https://doi.org/10.3390/w7052358, 2015.

Wagener, T., Sivapalan, M., Troch, P., and Woods, R.: Catchment classification and hydrologic similarity, Geography Compass, 1, 901–931, https://doi.org/10.1111/j.1749-8198.2007.00039.x, 2007.

Westerberg, I. K., Guerrero, J.-L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. E., and Xu, C.-Y.: Calibration of hydrological models using flow-duration curves, Hydrol. Earth Syst. Sci., 15, 2205–2227, https://doi.org/10.5194/hess-15-2205-2011, 2011.

Westerberg, I. K., Wagener, T., Coxon, G., McMillan, H. K., Castellarin, A., Montanari, A., and Freer, J.: Uncertainty in hydrological signatures for gauged and ungauged catchments, Water Resour. Res., 52, 1847–1865, https://doi.org/10.1002/2015WR017635, 2016.

Wolfe, W., Haugh, C., Webbers, A., and Diehl, T.: Preliminary conceptual models of the occurrence, fate, and transport of chlorinated solvents in karst regions of Tennessee, U.S. Geological Survey, Nashville, Tennessee, USA, Water Resources Investigations Report 97–4097, 88 pp., 1997.

Yadav, M., Wagener, T., and Gupta, H.: Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins, Adv. Water Resour., 30, 1756–1774, https://doi.org/10.1016/j.advwatres.2007.01.005, 2007.

Yilmaz, K. K., Gupta, H. V., and Wagener, T.: A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model, Water Resour. Res., 44, W09417, https://doi.org/10.1029/2007WR006716, 2008.

Paper III

In press in *Hydrological Sciences Journal*

Evaluating model performance: Towards a non-parametric variant of the Kling-Gupta efficiency

S. Pool[1], M. Vis[1], and J. Seibert[1, 2]

[1]University of Zurich, Department of Geography, Zurich, Switzerland

[2] Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment, Uppsala, Sweden

Corresponding author: Sandra Pool (sandra.pool@geo.uzh.ch)

Abstract: Goodness-of-fit measures are important for an objective evaluation of runoff model performance. The Kling-Gupta efficiency (R_{KG}), which has been introduced as an improvement of the widely used Nash-Sutcliffe efficiency, considers different types of model errors, namely the error in the mean, the variability and the dynamics. The calculation of R_{KG} is implicitly based on the assumptions of data linearity, data normality, and the absence of outliers. In this study, we propose a modification of R_{KG} as an efficiency measure comprising non-parametric components, i.e., the Spearman rank correlation and the normalized flow-duration curve. The performances of model simulations for 100 catchments using the new measure were compared to those obtained using R_{KG} based on a number of statistical metrics and hydrological signatures. The new measure resulted overall in better or comparable model performances, and thus, it was concluded that efficiency measures with non-parametric components provide a suitable alternative to commonly used measures.

Keywords: runoff modelling, calibration, non-parametric, multi-objective, Kling-Gupta efficiency

1 Introduction

Runoff models are important tools in hydrology. Their application requires some form of parameter estimation to ensure reliable discharge simulations for the catchment of interest. Parameter estimation is oftentimes based on comparing simulated and observed discharge using a goodness-of-fit measure, also called an objective function. The most widely used objective function in hydrological modelling is the model efficiency (Nash and Sutcliffe 1970), which is based on the mean squared error. The mean squared

error between simulated and observed discharge can be decomposed into the three components mean, variability, and dynamics (Murphy 1988, Gupta et al. 2009). Estimating model parameters by optimizing the mean squared error is critical in two ways. Gupta et al. (2009) demonstrated that a high model performance for discharge dynamics is inevitably related to an underestimation of discharge variability and that the importance of discharge volume in model calibration depends on a catchment's discharge variability. This motivated them to suggest an objective function (the so called Kling-Gupta model efficiency, RKG), which is based on an improved combination of the three diagnostically meaningful components of the mean squared error.

The Kling-Gupta model efficiency is in line with the paradigm of using multiple objectives for model calibration with the aim to prevent an overfitting of model parameters to a particular hydrograph aspect (some early studies are Lindström 1997, Gupta et al. 1998, Boyle 2000, Madsen 2003). Taking into account multiple objectives can reduce simulation uncertainties and provides more reliable predictions given that the individual objectives are uncorrelated (Efstratiadis and Koutsoyiannis 2010). Multi-objective functions were originally mostly composed of purely statistical metrics, such as the root mean squared error of low, high or peak flows (see review of Efstratiadis and Koutsoyiannis 2010). In more recent years, hydrological signatures were applied as multi-objective functions (Yilmaz et al. 2008, Hingray et al. 2010, Euser et al. 2013, Zhang et al. 2016, Kiesel et al. 2017, Shafii et al. 2017) with the aim of focusing model calibration on relevant hydrograph aspects or major catchment functions. The term multi-objective function can also refer to multiple variables or multiple sites within a catchment (Madsen 2003). In this study, however, we used only discharge time series for calibration.

The calculation of R_{KG} is implicitly based on the assumptions of data linearity and normality, as well as the absence of outliers. However, discharge time series and model simulation errors are known to be highly skewed, which violates the implicit assumptions underlying R_{KG}. The aim of this study was therefore to make a step towards using non-parametric efficiency measures by reformulating the variability and the correlation term of R_{KG} in a non-parametric form. For a non-parametric alternative to the standard deviation, we decided to use the flow-duration curve (FDC). The FDC describes the relationship between the frequency and magnitude of streamflow and is an indicator of flow variability across all flow magnitudes of a catchment (Vogel and Fennessey 1995), whereas the standard deviation is, in case of non-normally distributed data, only a metric for the variability of flows around the mean flow. Since catchment characteristics such as flashiness or baseflow can be linked to specific segments of the FDC, it has become a widely used signature for model calibration (Yilmaz et al. 2008, Westerberg et al. 2011, Pokhrel et al. 2012, Euser et al. 2013, Pfannerstill et al. 2014, Garcia et al. 2017). As proposed by Legates and McCabe (1999), we used the Spearman rank correlation to describe discharge dynamics instead of the Pearson

correlation coefficient as is used by R_{KG}. Spearman rank correlation is less sensitive to extreme values in a time series than Pearson correlation and is therefore less prone to artificially high correlation values, leading to a more robust characterization of the correlation (Legates and McCabe 1999, Krause et al. 2005). Spearman rank correlation is, just as the Pearson correlation, insensitive to additive and proportional differences between simulated and observed discharge (Legates and McCabe 1999), which stresses the importance of the volume-term in R_{KG}. To our knowledge, the Spearman rank correlation has only been used in a limited number of calibration studies (Vis et al. 2015 or Seibert and Vis 2016).

In this study, we propose a modification of R_{KG} towards a non-parametric calibration criterion (R_{NP}) that is composed of the mean discharge, the FDC, and the Spearman rank correlation. Model calibrations with R_{KG}, R_{NP}, and different combinations of their mean, variability, and dynamic components were evaluated by comparing the model performance for a number of selected hydrograph aspects. The goal was to evaluate the potential of non-parametric formulations of goodness-of-fit measures for runoff model calibrations aiming at multiple hydrograph aspects.

2 Data and methods

2.1 Study area

This study was based on model applications in 100 catchments located across the contiguous United States (Fig. 1). The catchments are a subset of the Newman et al. (2015) data set and were selected by stratified random sampling from the drainage area of the major river regions proportional to the number of gauged catchments in these river regions. The Newman et al. (2015) data set provides daily temperature, precipitation and discharge data along with catchment outline information for over 600 catchments in the United States with minimal human disturbance. Monthly potential evaporation was estimated using the Priestley-Taylor equation for which the required input data was as well extracted from the Newman et al. (2015) data set. The catchment areas range from 10 km2 to 12 630 km2 with a median of 340 km2. Mean catchment elevations are between 25 m a.s.l. and 3355 m a.s.l. Annual precipitation sums vary from 240 mm to 3070 mm, of which more than 15 % falls as snow in a third of the catchments (based on the time period from 1990 to 2009). From the study catchments 43 % can be classified as humid, 40 % as temperate and 17 % as arid (classification according to Coopersmith et al. 2014). The wide range of catchment areas and hydroclimatic conditions of the selected catchments ensured that a large variability in runoff processes were represented among the study catchments.

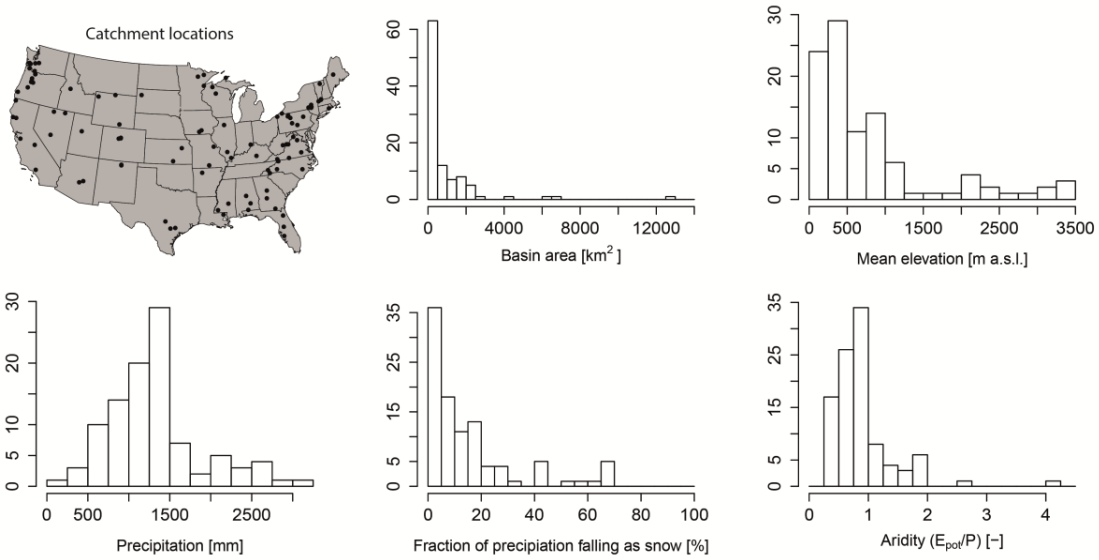


Figure 1. Locations and hydroclimatic characteristics of the 100 study catchments.

2.2 The HBV runoff model

The HBV runoff model (Hydrologiska Byråns Vattenbalansavdelning, Bergström, 1976, Lindström et al., 1997) in the version of HBV-light (Seibert and Vis, 2012) was used to test the influence of the different objective functions on simulated runoff. The runoff model has been successfully applied in many different hydroclimates (e.g. Häggström et al., 1990, Lidén and Harlin, 2000, Perrin et al., 2001, Beck et al., 2016, Seibert and Vis, 2016). The HBV model is a bucket-type runoff model with a conceptual representation of runoff processes at the catchment scale. The model consists of four routines representing snow, soil water, groundwater and stream network routing. Daily temperature and precipitation are input to the snow routine, where snow accumulation and melt are calculated with a degree-day method. Snowmelt and rainfall supply the soil moisture storage from which, together with monthly potential evaporation, the actual evaporation and the groundwater recharge is computed. Groundwater storage is represented by a shallow and a deep reservoir from which the fast runoff response, intermediate runoff response and baseflow are calculated. These three runoff components are summed and transformed by a triangular weighting function to simulate the hydrograph at the catchment outlet.

The HBV model was applied in a semi-distributed way by dividing the catchment into elevation bands of 200 m with separate computations for the snow and soil routines. Temperature and precipitation input to the elevation bands was calculated with a lapse rate of -0.6 °C per 100 m and 10 % per 100 m respectively. Potential evaporation was assumed to be uniform over the entire catchment. Elevation bands were determined using SRTM elevation data (Shuttle Radar Topography Mission, Jarvis et al., 2008).

2.3 Model calibration criteria

In this study, multiple model calibration criteria were defined that are based on the decomposition of the mean squared error into the three aspects, mean (β), variability (α) and dynamics (i.e., correlation r; Murphy 1988, Gupta et al. 2009).

The three terms β, α, and r were first calculated as originally proposed by Gupta et al. (2009; Eq. 1-3). The bias between simulated (*sim*) and observed (*obs*) mean discharge μ and the bias between simulated and observed standard deviation σ was used to compute β and α_{KG}, respectively. The Pearson correlation between observed and simulated discharge time series Q with length n was used as indicator for discharge dynamics (r_p). Together the three parametric components β, α_{KG}, and r_p were input to the Kling-Gupta efficiency R_{KG} (Eq. 4).

$$\beta = \frac{\mu_{sim}}{\mu_{obs}} \tag{1}$$

$$\alpha_{KG} = \frac{\sigma_{sim}}{\sigma_{obs}} \tag{2}$$

$$r_p = \frac{\sum_{i=1}^{n}(Q_{obs}(i) - \mu_{obs})(Q_{sim}(i) - \mu_{sim})}{\sqrt{\left(\sum_{i=1}^{n}(Q_{obs}(i) - \mu_{obs})^2\right)\left(\sum_{i=1}^{n}(Q_{sim}(i) - \mu_{sim})^2\right)}} \tag{3}$$

$$R_{KG} = 1 - \sqrt{(\beta - 1)^2 + (\alpha_{KG} - 1)^2 + (r_p - 1)^2} \tag{4}$$

To make a step towards a non-parametric variant of R_{KG}, the terms α and r were furthermore expressed in a non-parametric way. The non-parametric form of the discharge variability (α_{NP}) was built on the FDC. The FDC was normalized to remove the volume information and only keep the distribution signal. The absolute error was then computed between all ranked simulated and observed discharge values (Eq. 5; where I(k) and J(k) are the time steps when the kth largest flow occurs within the simulated and observed time series, respectively). For a non-parametric alternative to the correlation term the Spearman rank correlation (r_s) was calculated on the ranks of the observed (R_o) and simulated (R_s) discharge time series (Eq. 6). The combination of β, α_{NP}, and r_s into a single metric resulted in the partly non-parametric objective function R_{NP} (Eq. 7). An R-script with the calculation for R_{NP} is provided in the supplementary material.

$$\alpha_{NP} = 1 - \frac{1}{2}\sum_{k=1}^{n}\left|\frac{Q_{sim}(I(k))}{n\bar{Q}_{sim}} - \frac{Q_{obs}(J(k))}{n\bar{Q}_{obs}}\right| \tag{5}$$

$$r_s = \frac{\sum_{i=1}^{n}(R_{obs}(i) - \bar{R}_{obs})(R_{sim}(i) - \bar{R}_{sim})}{\sqrt{(\sum_{i=1}^{n}(R_{obs}(i) - \bar{R}_{obs})^2)(\sum_{i=1}^{n}(R_{sim}(i) - \bar{R}_{sim})^2)}} \quad (6)$$

$$R_{NP} = 1 - \sqrt{(\beta - 1)^2 + (\alpha_{NP} - 1)^2 + (r_s - 1)^2} \quad (7)$$

Overall, the components β, α, and r used in their parametric and non-parametric variants built the foundation for various one-, two-, and three- component objective functions used in this study (Fig. 2):

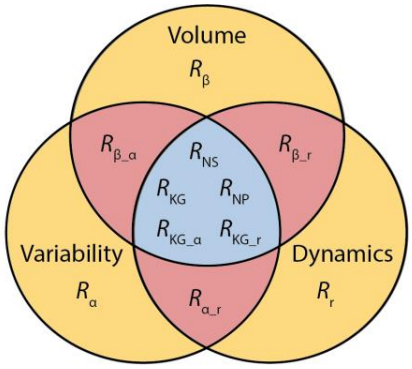
(1) One-component objective functions were defined so that each consisted of a single variable from R_{NP} (R_β, R_α, and R_r).

(2) Two-component objective functions consisted of two equally weighted variables from R_{NP} (R_{β_α}, R_{β_r}, and R_{α_r}).

(3) For the three-component objective functions we used β and both parametric and non-parametric variants of α and r. The Nash-Sutcliffe model efficiency (R_{NS}), the Kling-Gupta model efficiency (R_{KG}) and its non-parametric modification (R_{NP}) were assigned to this third group of objective functions. To complement R_{KG} and R_{NP}, two further objective functions were introduced where either α (R_{KG_α}) or r (R_{KG_r}) was modified to be non-parametric. These two versions were used to analyse the effect of each of the individual modifications that were made to R_{KG} in order to generate R_{NP}. Similar to R_{KG} and R_{NP}, the multiple components of R_{KG_α} and R_{KG_r} were combined using the Euclidean distance measure (Eq. 4 and 7).

2.4 Model calibration and evaluation

The HBV model was calibrated against the continuous daily discharge time series of the hydrological years 1990 to 1999 for each of the 100 study catchments. Model parameters were optimized within predefined parameter ranges using a genetic algorithm (Seibert 2000) and each of the 11 objective functions (Fig. 2). To consider parameter uncertainty, the parameter optimization was performed 100 times. The model calibration process resulted in an ensemble of 100 calibrated parameter sets for each catchment and objective function. These parameter sets were additionally used to simulate discharge for a validation period (1 October 2000 to 30 September 2009). For both calibration and validation a two year warming-up period was used to ensure suitable initial values for the state variables.



	Volume (β)	Variability (α)	Dynamcis (r)
R_{NS}	Mean	Standard deviation	Pearson
R_{KG}	Mean	Standard deviation	Pearson
R_{KG_α}	Mean	Flow-duration curve	Pearson
R_{KG_r}	Mean	Standard deviation	Spearman
R_{NP}	Mean	Flow-duration curve	Spearman
R_{β_α}	Mean	Flow-duration curve	-
R_{β_r}	Mean	-	Spearman
R_{α_r}	-	Flow-duration curve	Spearman
R_β	Mean	-	-
R_α	-	Flow-duration curve	-
R_r	-	-	Spearman

Figure 2. Objective functions used for model calibration. The basic components describing discharge volume (β), variability (α), and dynamics (r) are combined into eleven one-, two- or three-component objective functions.

Model simulations in calibration and validation were evaluated in three ways. First, we evaluated hydrograph uncertainty related to the use of different objective functions. The spread between the 100 simulated hydrographs of each catchment was used as information on how well an objective function constrains model parameters for a particular catchment. To evaluate this spread in simulated hydrographs, we computed the difference between the 0.05 and 0.95 quantile of the 100 simulated hydrographs at each time step in the calibration time period. The difference was then normalized by the observed discharge and evaluated for different discharge quantiles to see if simulation uncertainty differed for different flow conditions. Hydrograph uncertainty was evaluated for simulations based on the objective functions R_{KG}, R_{KG_β}, R_{KG_α}, and R_{NP}.

Second, the 100 simulated hydrographs of each catchment were evaluated in terms of R_{KG}, R_{NP}, and the (non-) parametric β, α, and r components. The further analysis was based on the median of the 100 efficiencies of each catchment. The median efficiencies from all catchments were used to compute cumulative distribution functions for each evaluation metric. Furthermore, we were interested to which extend R_{KG}, R_{NP}, and their (non-) parametric β, α, and r components are correlated with each other. Therefore, the Spearman rank correlation was calculated for different pairs among the three components (β, α, and r), $R_{KG,}$ and R_{NP} for simulations in the calibration period.

Lastly, model performance for simulations with each objective function was evaluated for three commonly

used statistical metrics and five hydrograph signatures that were not explicitly considered in calibration. These statistical metrics were the model efficiency calculated for peak flows (R_{NS_peak}), model efficiency calculated on logarithmic flow (R_{NS_logQ}), and R_{MARE}, a measure for low flows (1 minus the mean absolute relative error between observed and simulated flow). The chosen hydrograph signatures provide information on the major catchment functions by linking rainfall input to the flow response of a catchment (Yilmaz et al., 2008). The five signatures are the percent bias in runoff ratio (B_{rr}), the difference in watershed lag time (B_t), the percent bias in the high-flow segment of the FDC (B_{hf}), the slope of the mid-flow segment of the FDC (B_{FDC}), and the low-flow segment of the FDC (B_{lf}). The signatures were calculated according to Yilmaz et al. (2008), except for the watershed lag time, where only the difference in observed and simulated lag time, and not the percent bias, was calculated. Throughout this study, we always evaluated the absolute values of the percent bias or the absolute values of the difference between signatures. To statistically quantify the different effect of R_{KG} and R_{NP} on statistical metrics and signatures, we conducted a Wilcoxon signed-rank test (Wilcoxon 1945) using the median efficiency of each of the 100 study catchments.

3 Results

3.1 Evaluation of model simulations for R_{KG}, R_{NP} and their components

The calibrated model simulations in general reproduced the observed hydrographs reasonably well. The 100 simulated hydrographs resulting from calibration with R_{KG} and R_{NP} for two example catchments, one snow and one winter-rain dominated, indicated that independent of a catchment's runoff regime, the range of model simulations (0.05 to 0.95 quantile) is generally wider for simulations based on R_{KG} than R_{NP} (Fig. 3). This observation was confirmed by the results from all 100 catchments (Fig. 4). The difference in the range of simulated hydrographs was especially pronounced during recession periods and low-flow conditions. At exceptionally high peak flows (0.95 flow quantile), however, simulation uncertainty for calibrations with R_{NP} exceeded those of calibrations with R_{KG}. Simulation uncertainty resulted from the interplay between both the variability and the dynamics measure of R_{KG} and R_{NP} (Fig. 4). While variability and dynamics comparably influenced the simulation uncertainty for mean-flow conditions, their individual effect varied for low and high flows. During low-flow conditions, simulation uncertainty was most strongly influenced by the dynamics component of R_{KG} and R_{NP}, whereby the sensitivity of the Pearson correlation coefficient for high discharge values resulted in less confined simulations during low flows. At high-flow conditions, it was the described sensitivity of the Pearson correlation coefficient and the use of the FDC that reduced the range in simulated hydrographs.

The median model efficiencies of the 100 hydrograph simulations for each study catchment are presented in Fig. 5. As expected, model efficiencies for R_{KG} and R_{NP} decreased when moving from calibration (median R_{KG} 0.86 and median R_{NP} 0.85) to validation (median R_{KG} 0.77 and median R_{NP} 0.80). This decrease was more pronounced for the objective function the model was calibrated on. Interestingly, the discharge variability measured in terms of the standard deviation was underestimated for calibrations based on R_{NP} in 80 % of the catchments, as opposed to an almost equal fraction of catchments being under and overestimated for calibrations with R_{KG}. Hydrograph dynamics, measured in terms of the Pearson correlation coefficient, were well represented by simulations calibrated with R_{KG}. However, the same model calibrations performed relatively poorly in terms of the Spearman rank correlation coefficient stressing the stronger sensitivity of the Pearson correlation to discharge extremes than to discharge dynamics.

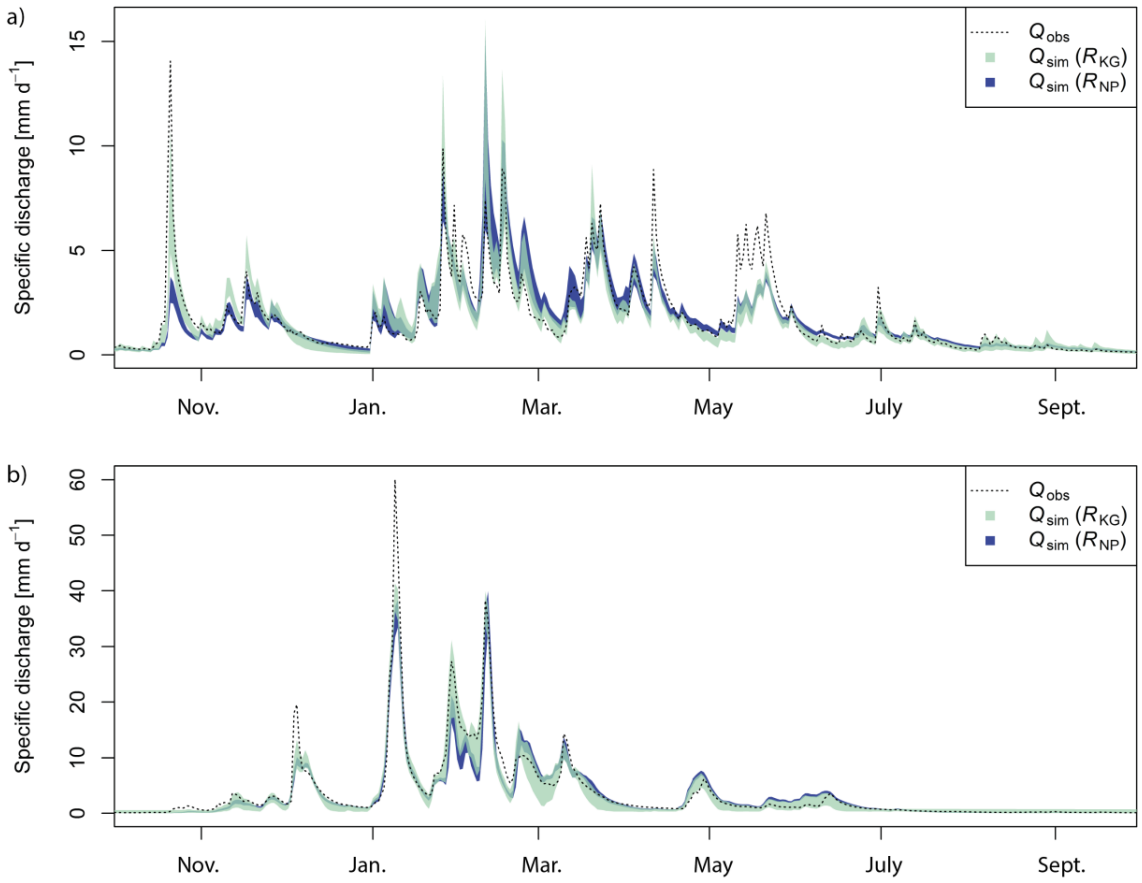


Figure 3. Observed and simulated hydrographs from model calibrations with R_{KG} and R_{NP} for a) a snow dominated catchment in the Northeast (USGS gauge id 01423000) and b) a winter-rain dominated catchment in the Northwest (USGS gauge id 14301000) of the United States. The range in hydrograph simulations indicates the 0.05 to 0.95 quantile of all 100 simulations.

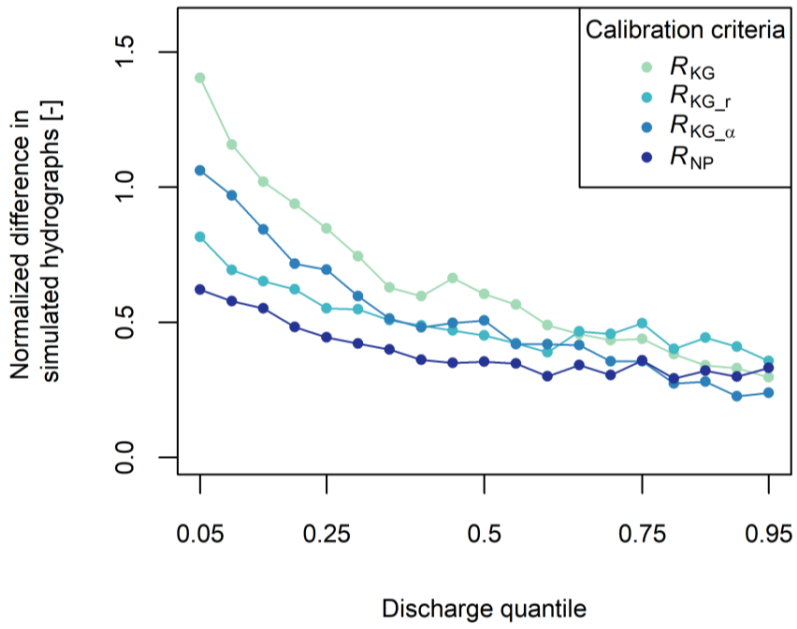


Figure 4. Hydrograph uncertainty for model calibrations with R_{KG}, R_{KG_r}, R_{KG_α}, and R_{NP}. For each catchment, uncertainty was calculated as the difference between the 0.05 and 0.95 quantile of the 100 hydrograph simulations at a particular point in time normalized by the observed discharge. Uncertainty was computed for various discharge quantiles. Here, the median uncertainty over all 100 study catchments is presented.

The individual components of a multi-objective function should ideally be uncorrelated to have a high information value for model calibration (Efstratiadis and Koutsoyiannis 2010). Table 1 shows that this requirement is only partly met for R_{KG} and R_{NP}. The rank correlation between r and β could be considered as weak, whereas it was strong between the r and α components and moderate to strong between α and β. The correlation between the multi-objective function (R_{KG} or R_{NP}) and its individual components (α, β, and r) is an indicator for the strength of their relation. Hydrograph dynamics were strongly related to the efficiency score of the multi-objective function, followed by the discharge variability component and the volume component, which was not necessarily well simulated when model efficiencies R_{KG} and R_{NP} were good.

Table 1. Spearman rank correlation coefficients for R_{KG} and R_{NP} and their three components in calibration. The correlation coefficients were calculated using the median values of all 100 catchments.

	R_{KG}				R_{NP}			
	R_{KG}	β	α_{KG}	r_p	R_{NP}	β	α_{NP}	r_s
$R_{KG/NP}$	1	0.42	0.58	0.97	1	0.27	0.71	0.97
β		1	0.61	0.3		1	0.34	0.21
$\alpha_{KG/NP}$			1	0.48			1	0.59
$r_{p/s}$				1				1

3.2 Effect of a stepwise modification of R_{KG} to R_{NP} on statistical metrics and hydrological signatures

The stepwise modification of the variability and correlation components of R_{KG} gives an indication of their individual effect on the model calibration with R_{NP} (Fig. 6). Statistical metrics (Fig. 6a) measuring model performance related to the magnitude and timing of high flows (R_{NS} and R_{NS_peak}) were better simulated with R_{KG} than R_{NP}. Adapting the variability component of R_{KG} by introducing the FDC led to negligible changes in model performance, whereas the replacement of the Pearson correlation by the Spearman rank correlation clearly impaired the timing and magnitudes of high flows. In contrast, the non-parametric variants of the variability and correlation components strongly improved the model performance for low-flow measures (R_{NS_logQ} and R_{MARE}) with the highest positive effect when changing both components simultaneously (R_{NP}). Similar effects as described for the statistical metrics could be observed for the high and low flow related hydrograph signatures (Fig. 6b). The two signatures runoff ratio and watershed lag time were not much affected by changes in the variability and correlation components.

Ranking the objective functions R_{KG}, R_{KG_r}, R_{KG_α}, and R_{NP} (Fig. 6c) according to their model performance provides a generalized picture of their effect on various hydrograph characteristics. The ranking highlights that the introduction of the non-parametric variant of the variability component (R_{KG_α}) often resulted in better simulations in comparison to R_{KG}. The non-parametric variant of R_{KG} could be considered as a valuable alternative for R_{KG}, unless timing and magnitude of high flows were of major importance.

The Wilcoxon signed-rank test revealed that the model efficiency for statistical metrics and signatures (Fig. 6) differed significantly for calibrations with R_{KG} and R_{NP}, except for the signatures B_{rr}, B_{hf}, and B_{lf}.

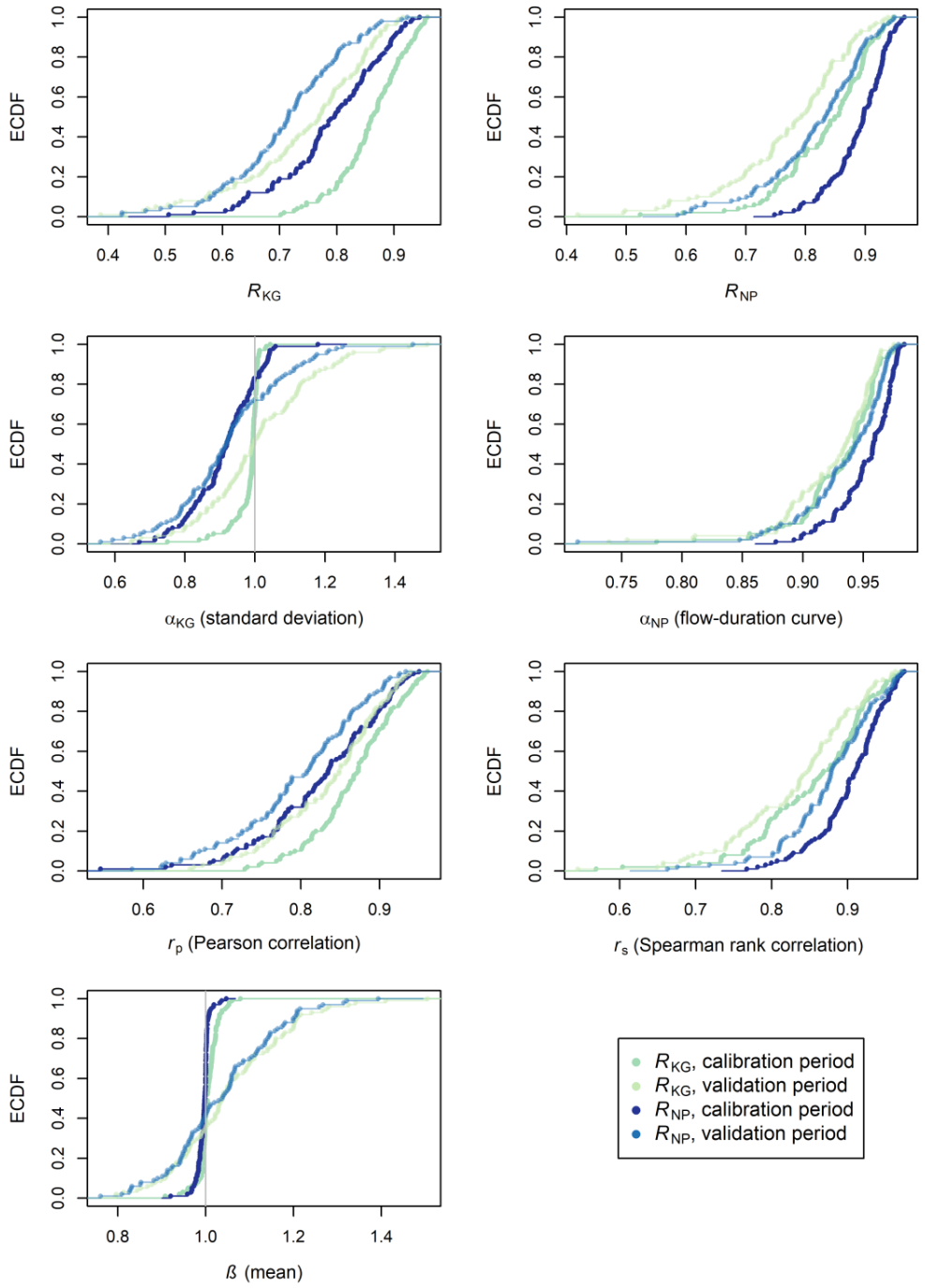


Figure 5. Model efficiencies (R_{KG}, R_{NP}, α_{KG}, α_{NP}, r_p, r_s, and β) in calibration and validation for model calibrations with R_{KG} and R_{NP}. Empirical cumulative distribution curves (ECDF) consist of the median model efficiency for each of the 100 study catchments.

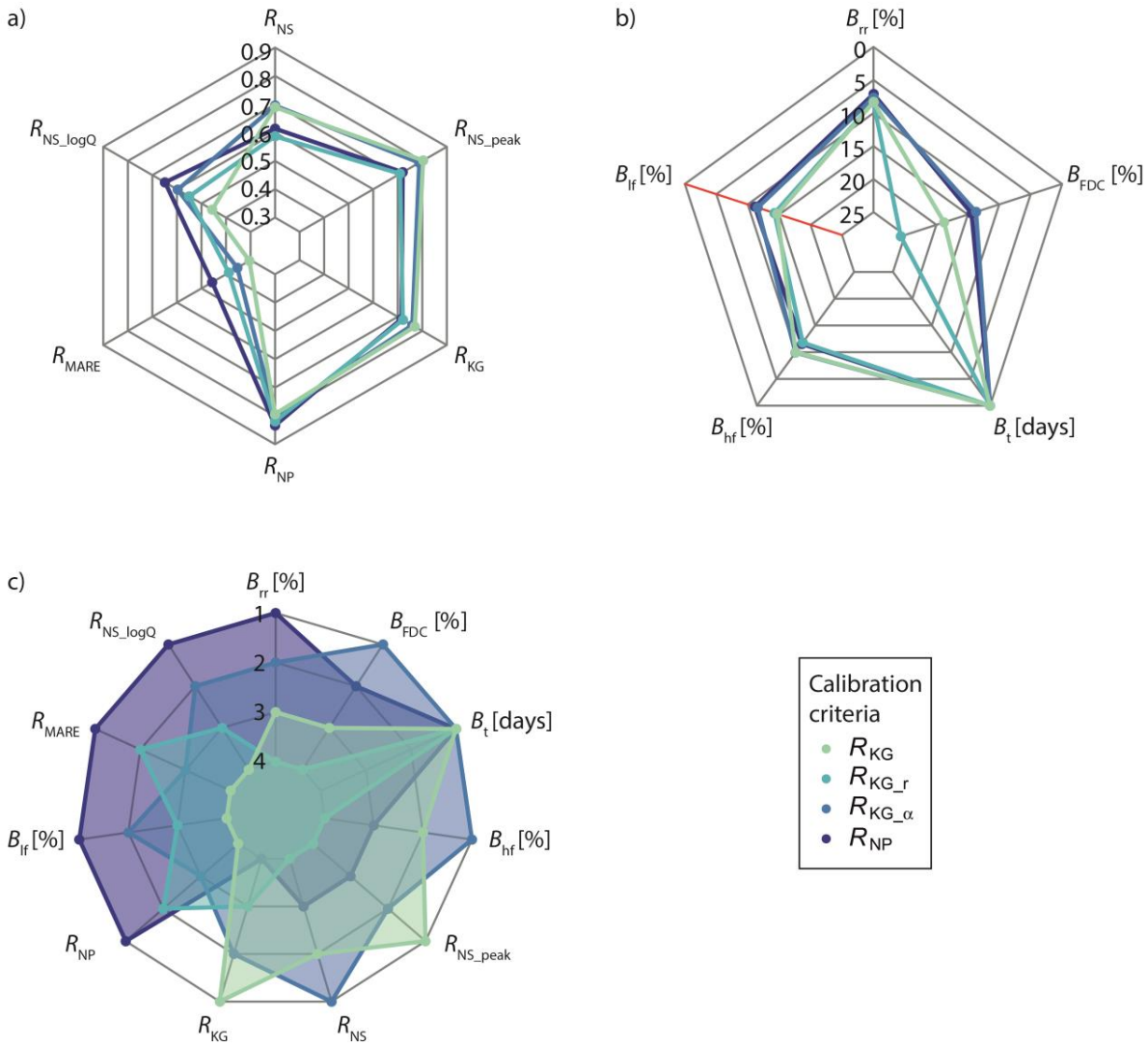


Figure 6. Model efficiencies in validation for model calibrations with R_{KG}, R_{KG_r}, R_{KG_α}, and R_{NP}. Calibration criteria are evaluated in terms of a) statistical metrics and b) hydrological signatures (note that the axis for B_{lf} is scaled by a factor of five meaning that percent bias is five times higher than indicated). Each calibration criterion is ranked according to its performance for statistical metrics and hydrological signatures in c). Results are presented for the median efficiency of all 100 study catchments.

3.3 Effect of the number of components on statistical metrics and signatures

Figures 7 and 8 present the results for model calibrations with nine objective functions consisting of a varying number of components for all catchments. For most statistical metrics and hydrograph signatures performance increased with an increasing number of components. Especially the loss of information on dynamics (by excluding the correlation component) negatively affected model performance in the two-component objective function. In the case of the one-component objective functions hydrograph dynamics were most important for model calibration, followed by the information on discharge variability. Model calibrations on volume only resulted in the poorest model simulations consistently throughout all evaluation metrics. Altogether, these results indicate that capturing all three components, i.e., discharge volume, variability, and dynamics of a catchment, is important for simulations aiming at multiple aspects of the hydrograph.

There are, of course, exceptions that do not follow the general observation made above. For example, the slope of the FDC (B_{FDC}) was best simulated when the α component, expressed in terms of the FDC, had a relatively high weight in calibration which was not the case for calibrations with a three-component objective function. Another exception is the percent bias in runoff ratio (B_{rr}) for which it was more essential to include a volume metric in the multi-objective function than to consider discharge dynamics. Lastly, discharge dynamics were less important than flow variability for simulating the high-flow segment of the FDC (B_{hf}) most likely because the timing is not of major relevance for that signature.

4 Discussion

The use of non-parametric goodness-of-fit measures is still a relatively new approach to model calibration. A comparison of various hydrograph characteristics resulting from calibrations with a partly non-parametric formulation of the popular Kling-Gupta efficiency (R_{NP}) and its original formulation (R_{KG}) demonstrated the potential of calibration criteria with non-parametric components. Overall, R_{NP} proved to be a valuable alternative for R_{KG}. It resulted in more confined hydrograph simulations (Figs. 3 and 4) and, except for high-flow metrics, in comparable or improved model performance for many of the statistical metrics and signatures (Fig. 6). Altogether, the flow-duration curve positively affected parameter selection, whereas the Spearman rank correlation had a varied effect on hydrograph simulations (Figs. 4 and 6).

More specifically, the use of the normalized FDC instead of the standard deviation had a positive effect on hydrograph simulations for all evaluated performance criteria. This favourable effect is encouraging although it may not be surprising. Unlike the standard deviation, which is a measure of discharge variability around the mean flow, the FDC contains information about the distribution of discharge over the full range

of magnitudes (Vogel and Fennessey 1995). It therefore supports the model calibration with more information on discharge variability than the standard deviation.

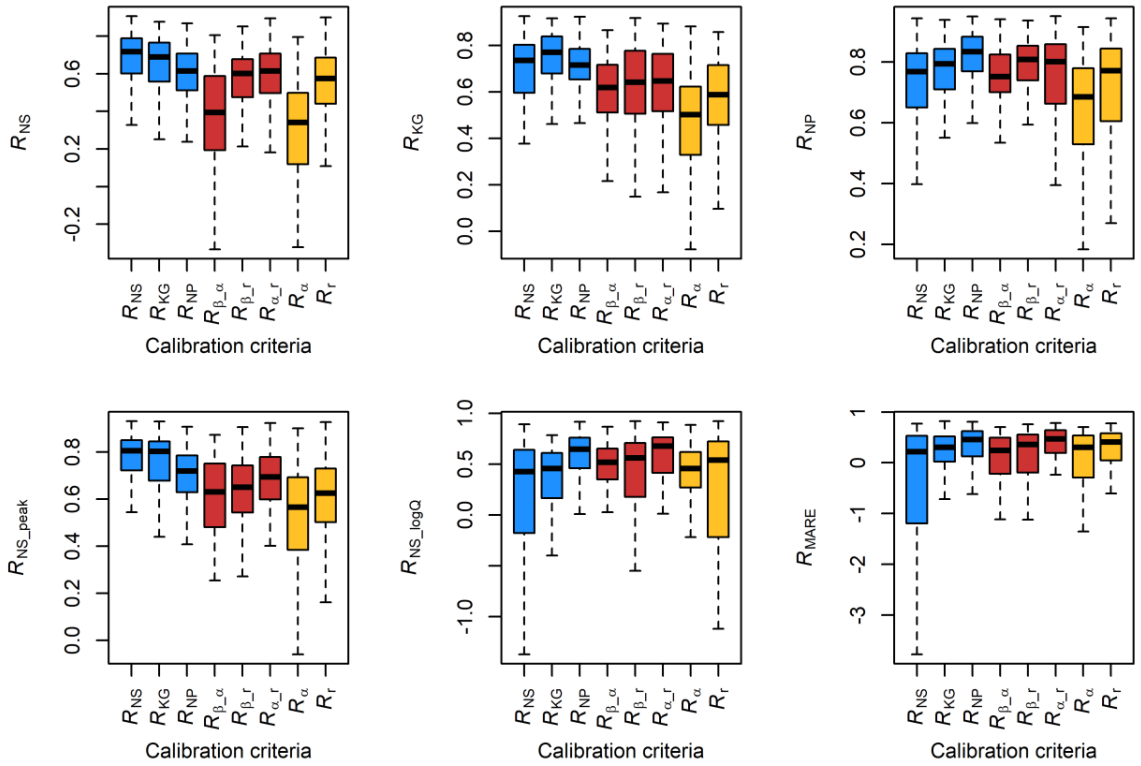


Figure 7. Model efficiencies (R_{NS}, R_{KG}, R_{NP}, R_{NS_peak}, R_{NS_logQ}, and R_{MARE}) in validation for model calibrations with three-, two- or one- component calibration criteria. Boxplots consist of the median model efficiency for each of the 100 study catchments. Note the different scales of the y-axis. Results for calibrations based on R_{β} are not displayed as they were much poorer than for all other calibrations. Median efficiencies for calibration with R_{β} were -0.50, -0.28, -0.14, 0.43, -1.13, and -2.55 for R_{NS}, R_{KG}, R_{NP}, R_{NS_peak}, R_{NS_logQ}, and R_{MARE} respectively.

A non-parametric formulation for discharge dynamics was especially valuable for simulating mean and low-flow conditions of a catchment as opposed to exceptionally high flow volumes or the timing of peak flows, which were better simulated when the Pearson correlation coefficient was used for model calibration. The sensitivity of the Pearson correlation coefficient to high discharge magnitudes (Legates and McCabe 1999, Krause et al. 2005) might seem beneficial for certain hydrograph aspects, but makes calibration sensitive to potential rating curve uncertainties at high flows. Given that model calibrations with R_{KG} (and therefore with the Pearson correlation coefficient) did not necessarily end in high Spearman rank

correlations questions the current predominant use of the Pearson correlation for describing discharge dynamics. The loss of information usually attributed to the use of Spearman rank correlation can therefore be a desirable effect when aiming at evaluating dynamics aspects. As a consequence, for many modelling applications Spearman rank correlation probably results in a more realistic representation of the overall dynamics and magnitudes of a catchment's runoff response than the Pearson correlation.

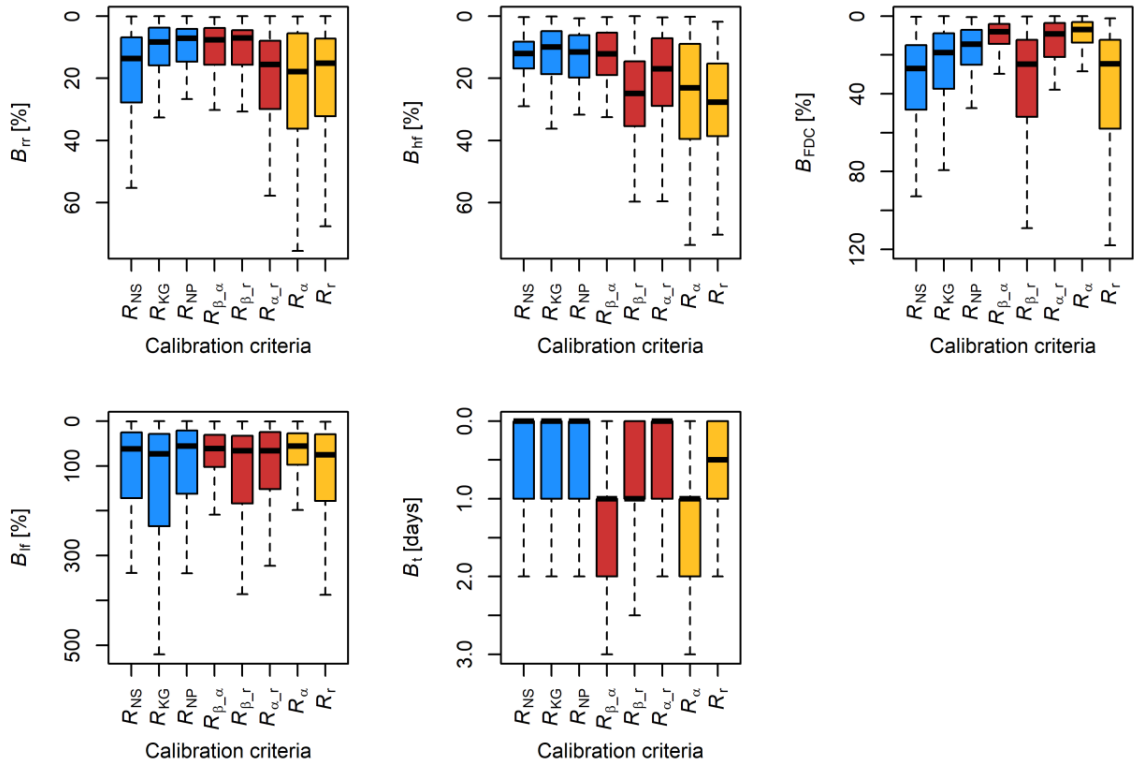


Figure 8. Model efficiencies ($B_{rr,}$ B_{hf}, B_{FDC}, B_{lf}, and B_t) in validation for model calibrations with three-, two- or one-component calibration criteria. Boxplots consist of the median model efficiency for each of the 100 study catchments. Note the different scales of the y-axis. Results for calibrations based on R_β are not displayed as they were much poorer than for all other calibrations. Median efficiencies for calibration with R_β were 96.0 %, 28.2 %, 62.7 %, 87.9 %, and 1.8 days for $B_{rr,}$ B_{hf}, B_{FDC}, B_{lf}, and B_t respectively.

Although the goal of the proposed modification of R_{KG} was to make a step towards non-parametric calibration criteria, we decided to use the mean instead of the median, which would have been the non-parametric alternative, to describe discharge volumes for two main reasons. First, information on the total discharge volume in a hydrological year is essential to close the water balance during model calibration, i.e., to constrain model parameters and ensure a correct simulation of actual evapotranspiration. For skewed

distributions, such as those of discharge time series, the median can deviate largely from the mean and a good model fit could have been achieved without closing the water balance. Second, the median discharge of semi-arid and arid catchments with prolonged dry periods might be zero, which would result in numerical problems when computing the ratios of simulated and observed values.

Mean discharge, normalized FDC, and Spearman rank correlation all provide unique information for model calibration that is not represented by one of the other criteria. As a consequence, using all three components for model calibration (R_{NP}) resulted in a better overall model performance than using a subset of the three components. These results are consistent with the observation that more robust hydrograph simulations are achieved with multi-objective model calibration (e.g., Lindström 1997, Gupta et al. 1998, Boyle 2000, Madsen 2003, Efstratiadis and Koutsoyiannis 2010). Since mean discharge, normalized FDC, and Spearman rank correlation represent different hydrograph characteristics, it was to some extent surprising to see the moderate to strong correlation between them. One explanation for this observation is that discharge is often closely related to precipitation input. Especially in humid catchments, discharge volume and variability are reasonably modelled as long as hydrograph dynamics are well captured by the runoff model (Seibert and Vis 2016). A correlation between efficiency criteria can to some degree be desirable as it inhibits solutions where only a single hydrograph aspect is well simulated while others are poorly represented.

By selecting objective functions for the evaluation of runoff models, we implicitly make assumptions about the statistical nature of discharge data and model simulation errors. These assumptions can be that a discharge time series is normally distributed or does not include any outliers. However, such assumptions are often violated when working with real data. We therefore argue that from a conceptual point of view it is desirable to use non-parametric formulations of objective functions requiring weaker assumptions that are more likely met by observed and simulated discharge data. From a results perspective, we demonstrated that good results can be achieved when using a multi-objective function with non-parametric components to calibrate a model for multiple hydrograph aspects. Our results can be considered as relatively robust given that modelling results were based on 100 catchments with long modelling time series and which represent a large variety of hydroclimates. In practice, modellers often use log-transformed discharge to put less emphasis on high flows. This approach should be avoided for computing R_{KG}, because using log-transformed discharge would result in R_{KG} values that are sensitive to discharge close to zero and that are dependent on the chosen flow unit (Santos et al. 2018). Therefore, using R_{NP} could provide a valuable alternative in cases where one otherwise would use log-transformed flows.

5 Conclusion

In this study, we propose a modified variant of the Kling-Gupta efficiency towards a non-parametric calibration criterion for hydrological models. In this modified formulation discharge volume is described by the mean discharge, discharge variability is represented by the FDC, and discharge dynamics are expressed in terms of Spearman rank correlation. Given the conceptual advantages of non-parametric calibration criteria, the goal was to evaluate the potential and limits of such a goodness-of-fit measure for simulating multiple hydrograph aspects simultaneously. The proposed calibration approach was tested on 100 catchments across the contiguous United States, which span a large range of hydroclimatic conditions. From the evaluation of the simulated hydrographs on commonly used statistical metrics and signatures that represent various hydrograph aspects, the following conclusions can be drawn:

(1) The non-parametric modification of the Kling-Gupta efficiency generally resulted in better agreement between simulated and observed discharge than the original formulation, except when evaluating the magnitude and timing of high flows. The proposed non-parametric based multi-objective function can therefore be seen as a useful alternative to existing performance measures when aiming at acceptable simulations of multiple hydrograph aspects.

(2) The use of the FDC instead of the standard deviation to describe discharge variability positively affected all evaluated hydrograph aspects, which is likely due to the complete information on the discharge distribution contained in the FDC.

(3) The Spearman rank correlation generally improved simulations during mean and low-flow conditions compared to the Pearson correlation, which can be attributed to the insensitivity of the Spearman rank correlation to extreme values strengthening its characterisation of discharge dynamics.

(4) The combination of all three components of the mean squared error, namely discharge volume, variability and dynamics, in a single objective function generally resulted in simulations which represent multiple hydrograph aspects well. In contrast, model calibrations with a subset of the three components put emphasise on rather specific hydrograph aspects at the expense of a realistic representation of several hydrograph characteristics simultaneously.

6 Acknowledgements

This work was supported by the University of Zurich. Hydrometeorological data and catchment shapefiles were made available by Newman et al. (2015). SRTM elevation data was provided by Jarvis et al. (2008).

We thank Gilles Kratzer for the discussion about the definitions of parametric and non-parametric metrics. Alexander Gelfan and two anonymous reviewers are acknowledged for their valuable comments on an earlier draft of this manuscript.

7 References

Beck, H. E., van Dijk, A. I., de Roo, A., Miralles, D. G., McVicar, T. R., Schellekens, J., and Bruijnzeel, L. A., 2016. Global-scale regionalization of hydrologic model parameters. *Water Resources Research*, 52(5), 3599-3622.

Bergström, S., 1976. Development and application of a conceptual runoff model for Scandinavian catchments. Swedish Meteorological and Hydrological Institute, Norrköping, Sweden, No. RHO 7, 134 pp.

Boyle, D. P., Gupta, H. V., and Sorooshian, S., 2000. Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resources Research*, 36(12), 3663-3674.

Coopersmith, E. J., Minsker, B. S., Sivapalan M., 2014. Patterns of regional hydroclimatic shifts: An analysis of changing hydrologic regimes. *Water Resources Research*, 50, 1960– 1983.

Efstratiadis, A., and Koutsoyiannis, D., 2010. One decade of multi-objective calibration approaches in hydrological modelling: a review. *Hydrological Sciences Journal*, 55(1), 58-78.

Euser, T., Winsemius, H. C., Hrachowitz, M., Fenicia, F., Uhlenbrook, S., and Savenije, H. H. G., 2013. A framework to assess the realism of model structures using hydrological signatures. *Hydrology and Earth System Sciences*, 17(5), 1893-1912.

Garcia, F., Folton, N., and Oudin, L., 2017. Which objective function to calibrate rainfall–runoff models for low-flow index simulations? *Hydrological Sciences Journal*, 62(7), 1149-1166.

Gupta, H. V., Sorooshian, S., and Yapo, P. O., 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resources Research*, 34(4), 751-763.

Gupta, H. V., Kling, H., Yilmaz, K. K., and Martinez, G. F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, 377(1-2), 80-91.

Häggström, M., Lindström, G., Cobos, C., Martinez, J. R., Merlos, L., Alonzo, R. D., Castillo, G., Sirias, C., Miranda, D., Granados, J., Alfaro, R., Robles, E., Rodriguez, M. and Alfaro, R. I., 1990. Application of the HBV model for flood forecasting in six central American rivers. Swedish Meteorological and Hydrological Institute, Norrköping, Sweden, No. RHO 27, 14 pp.

Hingray, B., Schaefli, B., Mezghani, A., and Hamdi, Y., 2010. Signature-based model calibration for hydrological prediction in mesoscale Alpine catchments. *Hydrological Sciences Journal*, 55(6), 1002-1016.

Jarvis A., Reuter H.I., Nelson A., Guevara E., 2008. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m. Available at: http://srtm.csi.cgiar.org, last access: January 2016.

Kiesel, J., Guse, B., Pfannerstill, M., Kakouei, K., Jähnig, S. C., and Fohrer, N., 2017. Improving hydrological model optimization for riverine species. *Ecological Indicators*, 80, 376-385.

Krause, P., Boyle, D. P., and Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. *Advances in geosciences*, 5, 89-97.

Legates, D. R., and McCabe, G. J., 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. *Water resources research*, 35(1), 233-241.

Lidén, R., and Harlin, J., 2000. Analysis of conceptual rainfall–runoff modelling performance in different climates. *Journal of hydrology*, 238(3-4), 231-247.

Lindström, G., 1997. A simple automatic calibration routine for the HBV model. *Hydrology Research*, 28(3), 153-168.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, 201, 272–288.

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Advances in water resources*, 26(2), 205-216.

Murphy, A. H., 1988. Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly weather review*, 116(12), 2417-2424.

Nash, J. E., and Sutcliffe, J. V, 1970: River flow forecasting through conceptual models part I - A discussion of principles, *Journal of Hydrology*, 10(3), 282–290.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., Viger, R. J., Blodgett, D., Brekke, L., Arnold, J. R., Hopson, T., and Duan, Q., 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, 19, 209–223.

Perrin, C., Michel, C., and Andréassian, V., 2001. Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *Journal of Hydrology*, 242(3-4), 275-301.

Pfannerstill, M., Guse, B., and Fohrer, N., 2014. Smart low flow signature metrics for an improved overall performance evaluation of hydrological models. *Journal of hydrology*, 510, 447-458.

Pokhrel, P., Yilmaz, K. K., and Gupta, H. V., 2012. Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. *Journal of Hydrology*, 418, 49-60.

Santos, L., Thirel, G., and Perrin, C., 2018. Technical note: Pitfalls in using log-transformed flows within the KGE criterion, *Hydrology and Earth System Sciences*, in review.

Shafii, M., Basu, N., Craig, J. R., Schiff, S. L., and Van Cappellen, P., 2017. A diagnostic approach to constraining flow partitioning in hydrologic models using a multiobjective optimization framework. *Water Resources Research*, 53(4), 3279-3301.

Seibert, J., 2000. Multi-Criteria calibration of a conceptual runoff model using a genetic algorithm, *Hydrology and Earth System Sciences*, 4, 215–224.

Seibert, J. and Vis, M. J. P., 2012: Teaching hydrological modeling with a user-friendly catchment-runoff-model software package, *Hydrology and Earth System Sciences*, 16, 3315–3325.

Seibert, J., and Vis, M. J., 2016. How informative are stream level observations in different geographic regions?. *Hydrological Processes*, 30(14), 2498-2508.

Vis, M., Knight, R., Pool, S., Wolfe, W., and Seibert, J., 2015. Model calibration criteria for estimating ecological flow characteristics. *Water*, 7(5), 2358-2381.

Vogel, R. M., and Fennessey, N. M., 1995. Flow duration curves II: a review of applications in water resources planning. *JAWRA Journal of the American Water Resources Association*, 31(6), 1029-1039.

Westerberg, I. K., Guerrero, J. L., Younger, P. M., Beven, K. J., Seibert, J., Halldin, S., Freer, J. R., and Xu, C. Y., 2011. Calibration of hydrological models using flow-duration curves. *Hydrology and Earth System Sciences*, 15(7), 2205.

Wilcoxon, F., 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin, 1*(6), 80-83.

Yilmaz, Koray K., Hoshin V. Gupta, and Wagener, T., 2008. A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resources Research*, 44, 9.

Zhang, Y., Shao, Q., Zhang, S., Zhai, X., and She, D., 2016. Multi-metric calibration of hydrological model to capture overall flow regimes. *Journal of Hydrology*, 539, 525-538.

Paper IV

Journal of Hydrology 554 (2017) 613–622

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Research papers

Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration?

Sandra Pool [a,*], Daniel Viviroli [a], Jan Seibert [a,b]

[a] Department of Geography, University of Zurich, Zurich, Switzerland
[b] Department of Earth Sciences, Uppsala University, Uppsala, Sweden

ARTICLE INFO

Article history:
Received 3 April 2017
Received in revised form 18 July 2017
Accepted 19 September 2017
Available online 20 September 2017
This manuscript was handled by K. Georgakakos, Editor-in-Chief, with the assistance of Yasuto Tachikawa, Associate Editor

Keywords:
Runoff modelling
Hydrograph prediction
Flow-duration curve prediction
Ungauged catchment
Sampling strategy
Value of data

ABSTRACT

Applications of runoff models usually rely on long and continuous runoff time series for model calibration. However, many catchments around the world are ungauged and estimating runoff for these catchments is challenging. One approach is to perform a few runoff measurements in a previously fully ungauged catchment and to constrain a runoff model by these measurements. In this study we investigated the value of such individual runoff measurements when taken at strategic points in time for applying a bucket-type runoff model (HBV) in ungauged catchments. Based on the assumption that a limited number of runoff measurements can be taken, we sought the optimal sampling strategy (i.e. when to measure the streamflow) to obtain the most informative data for constraining the runoff model. We used twenty gauged catchments across the eastern US, made the assumption that these catchments were ungauged, and applied different runoff sampling strategies. All tested strategies consisted of twelve runoff measurements within one year and ranged from simply using monthly flow maxima to a more complex selection of observation times. In each case the twelve runoff measurements were used to select 100 best parameter sets using a Monte Carlo calibration approach. Runoff simulations using these 'informed' parameter sets were then evaluated for an independent validation period in terms of the Nash-Sutcliffe efficiency of the hydrograph and the mean absolute relative error of the flow-duration curve. Model performance measures were normalized by relating them to an upper and a lower benchmark representing a well-informed and an uninformed model calibration. The hydrographs were best simulated with strategies including high runoff magnitudes as opposed to the flow-duration curves that were generally better estimated with strategies that captured low and mean flows. The choice of a sampling strategy covering the full range of runoff magnitudes enabled hydrograph and flow-duration curve simulations close to a well-informed model calibration. The differences among such strategies covering the full range of runoff magnitudes were small indicating that the exact choice of a strategy might be less crucial. Our study corroborates the information value of a small number of strategically selected runoff measurements for simulating runoff with a bucket-type runoff model in almost ungauged catchments.

1. Introduction

Sustainable management of water resources and mitigation of natural hazards in ungauged catchments strongly rely on accurate and reliable runoff estimates often predicted by rainfall-runoff models (Sivapalan et al., 2003). Runoff models used in hydrology all consist of parameters representing different catchment characteristics. The effective values of these parameters cannot be measured directly, because of their conceptual meaning or incommensurability issues. As a consequence, parameter values need to be defined or adapted in a calibration process by comparing observed and simulated catchment runoff response (Beven, 2012). After a decade of research on prediction of runoff in ungauged basins (PUB), it still remains a considerable challenge to calibrate runoff models for data scarce catchments (Hrachowitz et al., 2013).

A variety of approaches have been developed to estimate model parameters for ungauged catchments. For example, regionalization methods were proposed that either estimate individual parameter values from regressions relating model parameters to catchment characteristics or that transfer entire parameter sets from gauged donor catchments to the ungauged target catchment based on

* Corresponding author.
E-mail address: sandra.pool@geo.uzh.ch (S. Pool).

https://doi.org/10.1016/j.jhydrol.2017.09.037
0022-1694/

proximity or similarity measures (see e.g. Parajka et al. (2013) for an extended discussion). Hydrograph predictions from regionalization could be improved given that a few runoff measurements were available to further constrain model parameters (Rojas-Serna et al., 2006; Drogue and Plasse, 2014; Viviroli and Seibert, 2015; Rojas-Serna et al., 2016). Some authors assumed that a short and intensive field campaign could be carried out in the catchment of interest to collect data for model calibration. They tested the value of combining runoff data and additional data such as groundwater dynamics (Freer et al., 2004; Juston et al., 2009; Seibert and McDonnell, 2013), soil moisture (Hughes et al., 2014) or hydrochemical tracers (Uhlenbrook and Sieber, 2005) for model calibration.

The PUB initiative determined the evaluation of the value of runoff data for model calibration as one of their main objectives (Sivapalan et al., 2003). This induced a series of studies exploring the minimum length of a runoff time series necessary to obtain robust model calibrations. First studies typically tested model sensitivity related to continuously measured runoff. Between two and eight years of runoff data were reported as minimum requirement for robust model parameterizations independent of the selected calibration period (Harlin, 1991; Yapo et al., 1996; Xia et al., 2004; Vrugt et al., 2006; Merz et al., 2009). While there is a general agreement that model performance tends to improve with an increased length of calibration data, much smaller data sets have been shown to be of comparable value as long continuous time series (McIntyre and Wheater, 2004; Perrin et al., 2007; Seibert and Beven, 2009, Singh and Bárdossy, 2012; Seibert and McDonnell, 2013; Melsen et al., 2014). Perrin et al. (2007) successfully calibrated a runoff model with 350 runoff measurements selected randomly from an almost forty year continuous runoff series. Seibert and Beven (2009) reported that approximately sixteen runoff measurements randomly picked within one hydrological year could already provide information for an acceptable model calibration. An alternative to randomly extracting measurements from a time series is the selection of runoff samples in a strategic manner. Seibert and Beven (2009) demonstrated that maximum flows or a combination of maximum and recession data contained more information than minimum or mean flows. Results from Seibert and McDonnell (2013) indicated that one fully gauged event or ten observations during different high flow situations had a similar information value as three months of continuously measured data. Extracting unusual events from a time series, Singh and Bárdossy (2012) achieved reliable model simulations with less than 10% of the data from a continuous time series. Moreover, event based sampling strategies resulted in better model performances than strategies with measurements at fixed time intervals (McIntyre and Wheater, 2004; Juston et al., 2009; Seibert and McDonnell, 2013). Model calibration with a limited number of runoff measurements performed best in relatively wet catchments (Perrin et al., 2007; Sun et al., 2017), which is a common observation in rainfall runoff modelling even when long continuous time series are available, or when runoff samples are selected during a wet period (Yapo et al., 1996; Vrugt et al., 2006; Kim and Kaluarachchi, 2009; Melsen et al., 2014; Correa et al., 2016). In addition, the consideration of hydrological variability and of hydrologically important processes was found to be essential for the calibration process and the resulting simulation uncertainty (Harlin, 1991; Vrugt et al., 2006; Konz and Seibert, 2010; Singh and Bárdossy, 2012).

The present study aimed at finding the most informative runoff measurements for calibrating a hydrologic model with a limited number of strategically selected runoff samples in order to accurately simulate the hydrograph and the flow-duration curve (FDC) in almost ungauged catchments. Based on data from twenty gauged catchments in the eastern US, which were treated as hypothetically poorly gauged catchments, we evaluated the following assumptions:

1) There is an optimal strategy to decide on when to measure runoff in an ungauged catchment to obtain the most informative data for constraining a runoff model.
2) The optimal strategy is generally valid, i.e., does not depend on the catchment or simulation evaluation criteria.
3) Runoff measurements chosen with an optimal sampling strategy are of comparable value as a long continuous runoff time series.

In our study we assume that measurements actually can be taken at these strategic points in time such as on the day with maximum flow during a month. In practice, this is obviously not possible as the runoff during a month is not known beforehand. However, our study gives an indication on how useful a certain strategy could be at best.

2. Data and methods

2.1. Study catchments and runoff model

This study was based on twenty catchments across the eastern US (Fig. 1). Catchment data was extracted from the freely available large scale dataset of Newman et al. (2015). The dataset with over 600 basins spread over the contiguous US includes catchments with only minimal human disturbances and complete thirty-year forcing and runoff data series. We selected twenty catchments that are similar in terms of wetness and precipitation seasonality, but different regarding the importance of snow related runoff processes. This small catchment sample can be considered as a relatively controlled subset of the large dataset with small hydroclimatic variation, but representing some of the most common runoff regime types in the US. The selected catchments (Table 1) vary in area from 148 to 2925 km^2 with steepest elevation gradients in or close to the Appalachian Mountains. Some catchments are to a large degree composed of wetlands and lakes account for up to 6% of the area of three of these catchments (C1, C2 and C20 in Table 1; Lehner and Döll, 2004). All catchments are humid and receive precipitation throughout the entire year. Snow processes dominate the runoff regime in northern latitudes where 10–28% of the annual precipitation falls as snow. The contribution of baseflow to runoff varies between the catchments from 23 to 69% indicating a large variation in runoff response characteristics.

Continuous daily runoff time series at the catchment outlets were simulated with a bucket-type runoff model, namely the HBV model (Hydrologiska Byråns Vattenbalansavdelning; Bergström, 1976; Lindström et al., 1997) in the version HBV-light (Seibert and Vis, 2012). The HBV model is forced with daily temperature and precipitation and monthly potential evaporation data. Hydrological processes are modelled with four model routines representing snow, soil water, groundwater and routing related processes. Snow accumulation and snowmelt are calculated in the snow routine using a degree-day method. Together with rainfall and potential evaporation, snowmelt is used to determine the actual evaporation and groundwater recharge in the soil routine. The groundwater routine consists of a shallow and a deep groundwater storage where the contribution of groundwater to peak runoff, intermediate runoff and baseflow is calculated. The routing routine transforms these three runoff components into the hydrograph at the catchment outlet by a triangular weighting function.

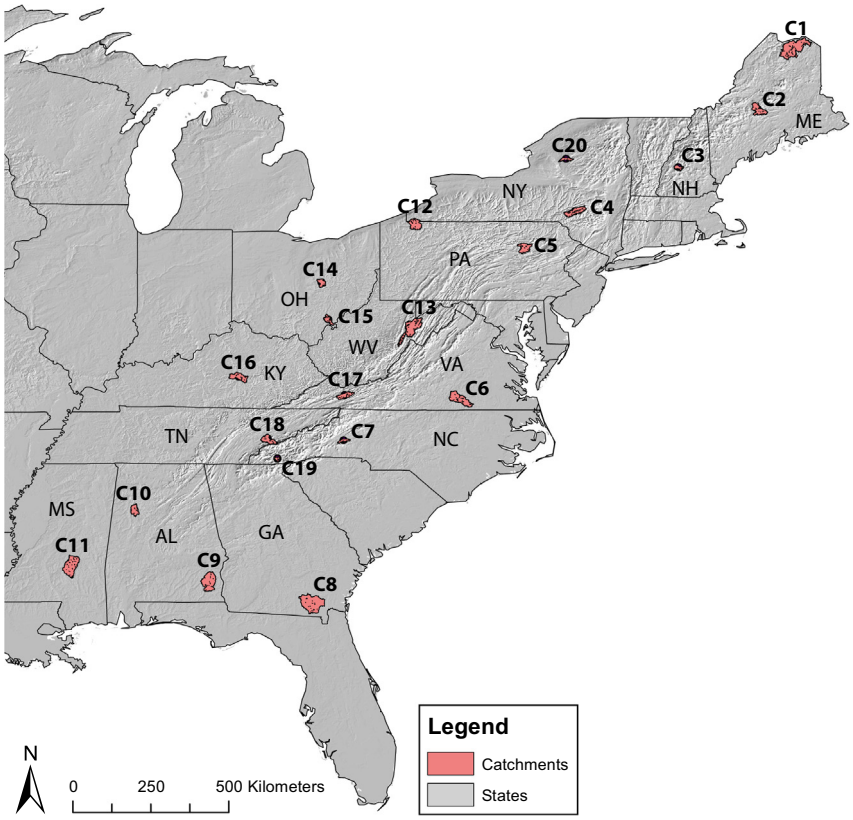


Fig. 1. Location of the twenty study catchments across the eastern US (catchment shapefiles from Newman et al. (2015); state boundaries and shaded relief from ESRI and U. S. Geological Survey (2011)).

Table 1
Information on the twenty study catchments. Snow [%]: percentage of annual precipitation falling as snow; precipitation seasonality: calculated according to Coopersmith et al. (2014), low seasonality for values ~<0.25; aridity index: ratio of sum of potential evaporation and sum of precipitation; runoff coefficient: ratio of runoff and sum of precipitation; baseflow [%]: percentage of runoff classified as baseflow, calculated based on the minimum runoff in fixed 5 day time intervals using the U.S. Geological Survey (2014) EflowStats R-package; wetland area [%]: percentage of catchment area covered by partial wetlands according to Lehner and Döll (2004).

ID	USGS station number and name	Area [km^2]	Mean elevation [m a.s.l.]	Snow [%]	Precipitation seasonality	Aridity index	Runoff coefficient	Baseflow [%]	Wetland area [%]
C1	01013500 Fish River near Fort Kent, ME	2260	379	27.6	0.17	0.63	0.54	68.9	92.2
C2	01031500 Piscataquis River near Dover-Foxcroft, ME	771	452	24.5	0.12	0.60	0.58	43.2	95.9
C3	01078000 Smith River near Bristol, NH	222	486	19.7	0.11	0.62	0.49	44.3	97.8
C4	01423000 West Branch Delaware River at Walton, NY	860	690	18.3	0.11	0.62	0.49	46.0	5.1
C5	01539000 Fishing Creek near Bloomsburg, PA	709	478	12.5	0.11	0.69	0.51	46.1	9.1
C6	02051500 Meherrin River near Lawrenceville, VA	1429	124	3.5	0.07	0.85	0.27	40.7	0.0
C7	02143000 Henry Fork near Henry River, NC	215	593	2.2	0.06	0.76	0.39	61.5	0.0
C8	02314500 Suwannee River at US 441 at Fargo, GA	2925	69	0.0	0.26	0.88	0.19	69.5	99.1
C9	02361000 Choctawhatchee River near Newton, AL	1776	127	0.0	0.16	0.82	0.31	52.5	0.0
C10	02464000 North River near Samantha, AL	577	157	0.9	0.12	0.70	0.37	29.6	0.0
C11	02472000 Leaf River near Collins, MS	1924	131	0.3	0.14	0.75	0.32	31.5	28.4
C12	03015500 Brokenstraw Creek at Youngsville, PA	831	486	16.3	0.14	0.63	0.54	40.2	21.4
C13	03069500 Cheat River near Parsons, WV	1869	984	16.4	0.11	0.61	0.60	36.2	21.6
C14	03144000 Wakatomika Creek near Frazeysburg, OH	362	308	7.4	0.13	0.84	0.36	36.6	0.0
C15	03159540 Shade River near Chester, OH	404	246	5.9	0.10	0.82	0.34	25.6	0.0
C16	03285000 Dix River near Danville, KY	823	349	3.5	0.10	0.77	0.40	23.0	0.0
C17	03488000 N F Holston River near Gate City, VA	572	976	6.8	0.11	0.81	0.38	46.1	0.0
C18	03498500 Little River near Maryville, TN	696	1141	2.9	0.11	0.64	0.41	51.8	0.0
C19	03500240 Cartoogechaye Creek near Franklin, NC	148	1121	2.4	0.09	0.55	0.45	68.1	0.0
C20	04256000 Independence River at Donnattsburg, NY	230	478	24.7	0.11	0.60	0.62	47.8	97.6

The HBV model allows runoff to be simulated in a semi-distributed way by disaggregating a catchment into elevation bands. We therefore split the catchments into elevation bands of 200 m using SRTM elevation data (Shuttle Radar Topography Mission; Jarvis et al., 2008). Temperature and precipitation data for each elevation band were interpolated with lapse rates of 0.6 °C per 100 m and 10% per 100 m, respectively. Potential evaporation was assumed to be uniform over all elevation bands and was calculated with the Priestley-Taylor equation.

2.2. Definition of sampling strategies

Sampling strategies were defined considering both existing hydrological knowledge from previous studies (see Section 1) and

practical aspects for the implementation of a runoff monitoring in the ungauged catchment of interest (Fig. 2). We defined a total of thirteen sampling strategies that were categorized as simple (*S*), intermediate (*I*) or complex (*C*) according to their hydrological background. For practical reasons it was interesting to examine sampling strategies with runoff samples at a fixed time interval (e.g. S_{DOM}). Runoff samples of event peaks or during low flow (e.g. S_{Max} or S_{Min}) could also be collected with relatively little effort as long as the exact timing was not crucial. From a hydrological point of view, strategies capturing runoff variability or dominant runoff processes could be promising. For example, the strategy $I_{Quantile}$ contains samples over the full range of runoff magnitudes, $C_{Max\_Min\_Wetness}$ takes into account the different runoff response of catchments after dry and wet periods or additional samples are taken during the snowmelt season with $C_{Max\_Snowmelt}$. All tested sampling strategies were restricted to twelve runoff samples within a single hydrological year (1st of October until 30th of September) that were extracted from the continuous runoff time series of each catchment. The decision to test the temporal distribution of runoff at twelve times within a year was chosen to represent a balance between a minimum number of measurements assumed to be necessary for model calibration and the practical limitations of measuring runoff at several times.

2.3. Modelling approach

The runoff model was calibrated for the twenty study catchments with a limited number of runoff samples. To run the model, twelve runoff samples selected from different hydrological years and the continuous precipitation and temperature data series were used in all cases. The data of fourteen hydrological years from 1983 to 1996 were used for independent model calibrations. A warm-up period of 2.75 years preceded each calibration period to ensure suitable initial values for the state variables. Model parameters of each calibration period were evaluated in an independent continuous validation time period from 1997 to 2010 in terms of how well the simulated runoff represented the observed hydrograph and the flow-duration curve. The two modelling time periods (1983–1996 and 1997–2010) were generally similar with respect to the yearly sum of precipitation, the yearly sum of runoff, the mean annual temperature and the percentage of precipitation falling as snow in each of the twenty study catchments (statistically evaluated using a non-parametric Mann-Whitney-*U* test). The detailed modelling steps were as follows:

1. 100,000 parameter sets were randomly generated within predefined parameter ranges (Table 2) and assuming a uniform parameter distribution.
2. The model was run for each parameter set. The simulated runoff was compared to the twelve observed runoff samples of each sampling strategy and calibration period. The objective functions used for comparison were the model efficiency (Nash and Sutcliffe, 1970) calculated directly on the runoff data (R_{eff}) and the model efficiency calculated on the log-transformed runoff data (R_{eff_logQ}). The 100 best parameter sets of each calibration period were retained for each strategy and objective function.

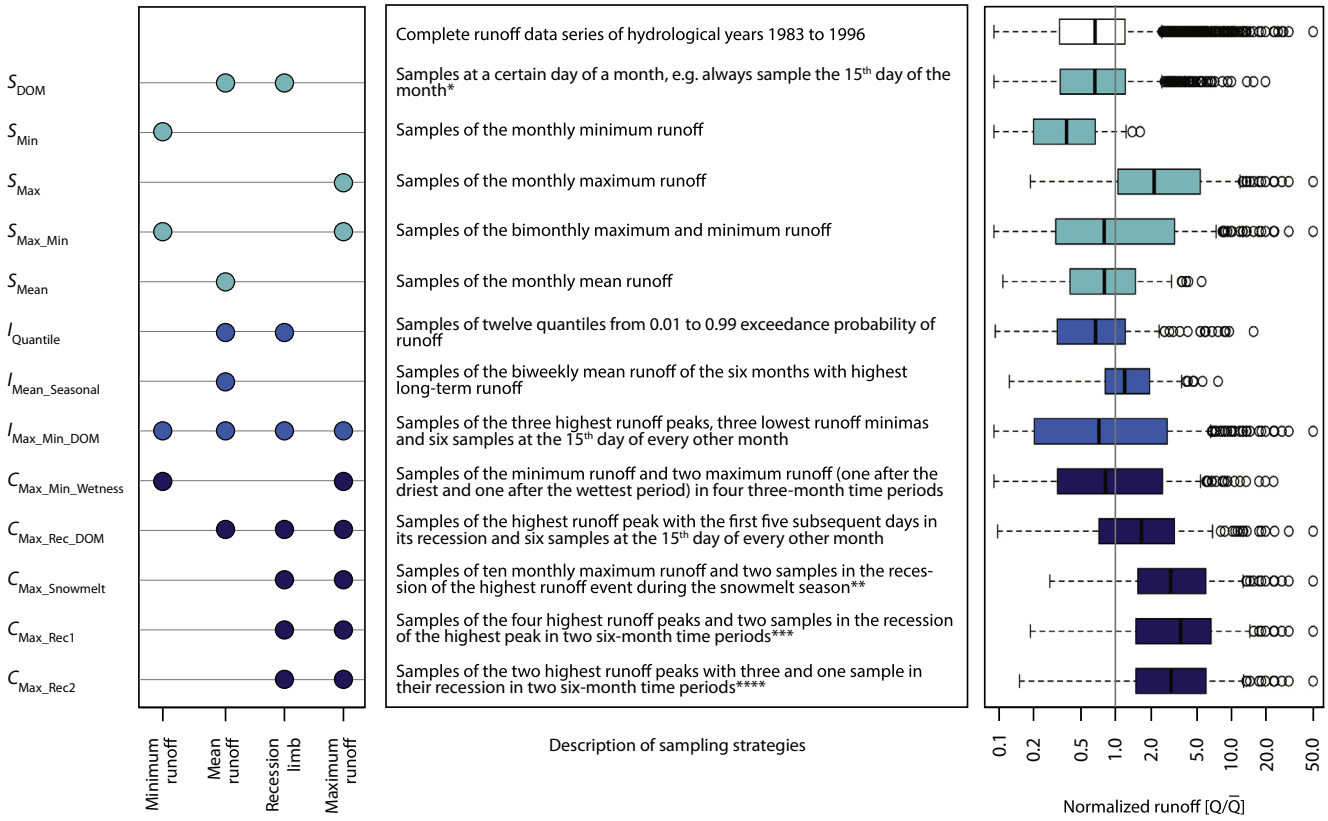


Fig. 2. Definition of the thirteen sampling strategies used for model calibration. Each sampling strategy consisted of twelve runoff samples. From left to right: abbreviation of sampling strategies, conceptual idea of runoff represented by strategies, description of strategies and normalized runoff magnitudes sampled with the strategies (normalized runoff corresponds to the sampled runoff Q divided by the mean catchment runoff $\bar{Q}$; data of catchment 17 (see Table 1) is shown). $^{*}S_{DOM}$: we tested the strategy with samples at the 1st, 5th, 10th, 15th, 20th and 25th day of the month and finally calculated the mean performance of all these six versions; $^{**}C_{Max\_Snowmelt}$: maximum runoff of the ten months with highest long-term runoff and recession samples taken at 80% and 60% of highest runoff peak in the snowmelt season (February to May); $^{***}C_{Max\_Rec1}$: recession samples taken at 80% and 40% of highest runoff peak; $^{****}C_{Max\_Rec2}$: recession samples taken at 80%, 60% and 40% of highest runoff peak and 80% of second highest runoff peak.

Table 2
Specification of HBV-light model parameters calibrated in this study according to Seibert and Vis (2012).

Parameter	Meaning	Unit	Minimum	Maximum
Snow routine				
TT	Threshold temperature	°C	−2	2.5
CFMAX	Degree-day factor	mm°C^{-1} d^{-1}	0.5	10
SFCF	Snowfall correction factor	–	0.5	1.2
SCR	Refreezing coefficient	–	0	0.1
CWH	Water holding capacity	–	0	0.2
Soil routine				
FC	Maximum soil moisture storage (SM)	mm	100	550
LP	Threshold for reduction of evaporation (SM/FC)	–	0.3	1
BETA	Shape coefficient	–	1	5
Groundwater routine				
PERC	Maximal flow from upper to lower box	mm d^{-1}	0	4
UZL	Maximal storage in the soil upper zone	mm	0	70
K0	Recession coefficient of fast response	d^{-1}	0.1	0.5
K1	Recession coefficient of intermediate response	d^{-1}	0.01	0.2
K2	Recession coefficient of baseflow	d^{-1}	0.00005	0.1
Routing routine				
MAXBAS	Routing, length of weighting function	d	1	5

3. The 100 best parameter sets were used to simulate runoff in the validation period. An ensemble mean hydrograph and ensemble mean FDC were calculated from the 100 runoff simulations. The ensemble mean hydrograph was evaluated in terms of R_{eff}. The ensemble mean FDC was evaluated by calculating the mean absolute relative error at 99 evaluation points of the FDC (R_{FDC}). The evaluation points were selected at equally spaced intervals of runoff volume between 0.1 and 0.99 exceedance probability, which is a similar approach to that suggested by Westerberg et al. (2011).

Model performance values in validation were normalized by relating them to an upper and a lower benchmark (Eq. (1)) as suggested by Girons Lopez and Seibert (2016). The upper benchmark represented the best possible model performance that could be achieved for a particular catchment. It was calculated with the simulation approach described above with the exception that the model was calibrated against the full continuous runoff time series of all fourteen years. While the upper benchmark parameter sets for the hydrograph were selected by applying R_{eff} or R_{eff_logQ}, R_{FDC} was used in the case of the FDC. The lower benchmark was calculated from 1000 randomly selected parameter sets and was a measure of how well the model would simulate runoff without any runoff information for a calibration. The identical normalization was applied for R_{eff} and R_{FDC} using the equation

$$R^* = \frac{R_{ss} - R_{lb}}{R_{ub} - R_{lb}} \quad (1)$$

with R^* as the normalized model performance (specifically R^*_{eff} and R^*_{FDC}), R_{ss} as the model performance based on the sampling strategy, R_{ub} as the model performance of the upper benchmark and R_{lb} as the model performance of the lower benchmark. Normalized performance values ranged from –inf to 1. A normalized performance of one indicates that model calibration with a particular sampling strategy was as good as a well-informed model calibration, whereas values below zero reveal that model calibration with a small number of strategically selected runoff measurements performs worse than simulations with random parameter sets.

Additionally, we evaluated the influence of the thirteen different sampling strategies for constraining model parameters. Since parameter values vary between catchments, we evaluated the range of parameter values, which had been calibrated based on a particular sampling strategy. Parameter ranges after calibration (0.05–0.95 quantile of all 100 parameter values) were normalized by their allowed range before calibration to make the different parameters comparable.

3. Results

When calibrated against the complete runoff time series, model performances were generally good for both the hydrograph (R_{eff}) and the FDC (R_{FDC}) (median R_{eff} 0.76 and median R_{FDC} 0.15; Fig. 3a and b, where the best possible model performance is 1.0 for the hydrograph and 0.0 for the FDC). As expected, model performances were poorer for simulations with a random parameterization (median R_{eff} 0.45 and median R_{FDC} 0.43). Model calibrations based on twelve runoff values selected by the different sampling strategies mostly resulted in performances between the two benchmarks. The hydrograph efficiency R_{eff} for all catchments and all strategies (Fig. 3a) ranged from 0.45 to 0.74 (median of 0.64) when parameter sets were selected based on R_{eff}. Calibrating the model with R_{eff_logQ} resulted in similar model performance for the hydrograph (R_{eff} from 0.48 to 0.74 with a median of 0.66) as calibrations with R_{eff}. Simulations of the FDC with a limited number of measurements (Fig. 3b) were considerably better when using the objective function R_{eff_logQ} instead of R_{eff}. Median R_{FDC} was 0.26 (range from 0.16 to 0.97) for calibrations with R_{eff_logQ} and 0.34 (range from 0.19 to 5.45) for calibrations with R_{eff}.

Model calibration with runoff data of a sampling strategy resulted in fourteen ensemble mean efficiencies for each catchment. The median of these fourteen values is an indicator of the information value of a particular strategy for model calibration. Ranking sampling strategies according to their median R^*_{eff} and R^*_{FDC} values revealed an interesting pattern with marked differences for the two evaluation criteria (Fig. 4a and b). The best ranked strategies for simulating the hydrograph (Fig. 4a) consisted of maximum runoff values mostly in combination with data in the recession of an event (e.g. $C_{Max\_Snowmelt}$). Strategies that combine maximum runoff with minimum runoff or runoff taken at a fixed time interval ranked in the middle (e.g. S_{Max_Min}). The poorest model performance was achieved by sampling minimum and mean runoff or by taking samples at a fixed time interval (e.g. S_{Min}). The described ranking pattern for the hydrograph was almost reversed when strategies were evaluated in terms of their information value for the FDC (Fig. 4b). The rank of each strategy was more consistent between the study catchments for the FDC than for the hydrograph. The differences in the ranking of strategies between catchments for the hydrograph simulation could partly be explained by catchment area and snowfall

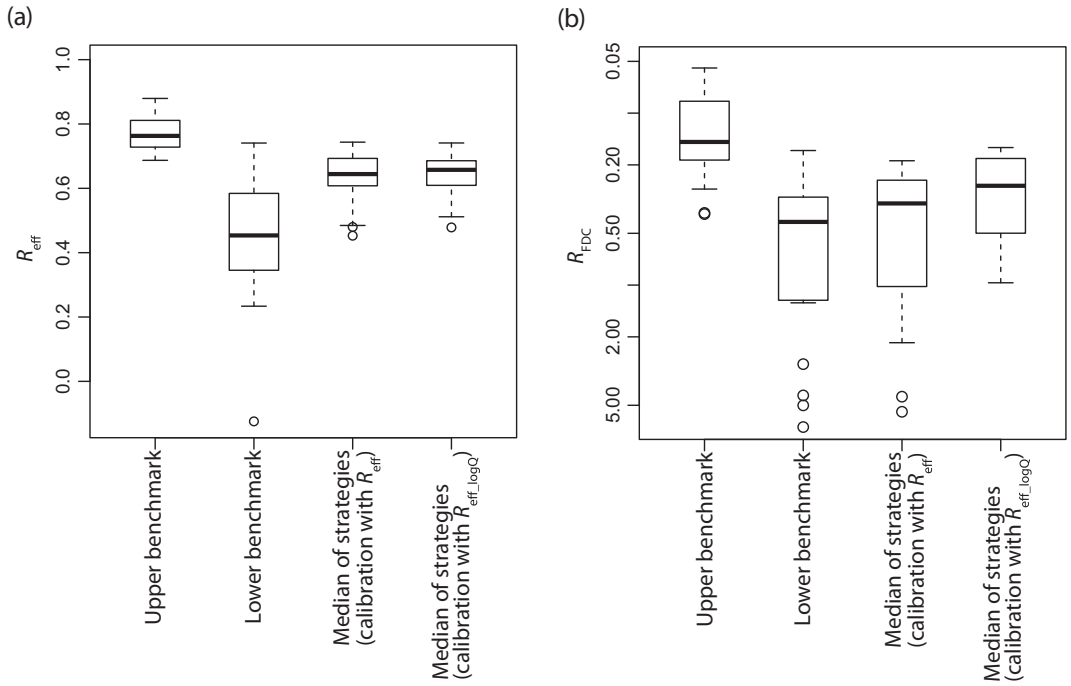


Fig. 3. Model performance for the twenty catchments as validated in terms of a) hydrograph efficiency R_{eff} and b) FDC efficiency R_{FDC} for model calibrations with the upper benchmark (continuous fourteen year calibration period), the lower benchmark (random generation of parameter sets) and the sampling strategies (twelve runoff samples) using either R_{eff} or R_{eff_logQ} as objective function. Best possible model performance is 1.0 for R_{eff} and 0.0 for R_{FDC}. Model performance related to the benchmarks was calculated as the median ensemble mean model performance of all calibration years for each catchment. Model performance of the sampling strategies is summarized by the median model performance of all strategies for each catchment. Strategy performance was calculated on the basis of the median ensemble mean performance of all calibration years.

ratio, whereby large catchments or small snow-dominated catchments tended to form clusters with a slightly different ranking of the sampling strategies. Other catchment characteristics such as mean elevation, precipitation seasonality, aridity, importance of baseflow or percentage of wetland area did not help to explain the mentioned variations. Not all strategies were more informative for model calibration than the lower benchmark with random parameter sets (Fig. 4). Especially catchments with a high model performance of the lower benchmark (R_{eff} >0.7), such as catchment C3, C9 and C11, had many sampling strategies with a negative normalized model performance for the hydrograph. Negative R^*_{FDC} values were most prominent in the low ranked sampling strategies ($I_{Mean\_Seasonal}$, C_{Max_Rec2}, $C_{Max\_Snowmelt}$, and C_{Max_Rec1}), suggesting that these strategies cannot be considered as an acceptable option for deciding on when to make runoff measurements in many catchments.

To evaluate the impact of using either R_{eff} or R_{eff_logQ} as objective function on the evaluation of the different sampling strategies, we focused on the median R^*_{eff} and median R^*_{FDC} values of a strategy over all catchments (Fig. 5a and b). Samples of maximum runoff were always crucial for a good hydrograph simulation, whereby the magnitude or timing of additional samples seemed to be of minor importance (e.g. S_{Max} or $C_{Max\_Rec\_Dom}$). R^*_{eff} values were between 0.52 and 0.72 for strategies containing high runoff values, independent of which of the two objective functions was applied in model calibration. In contrast, R^*_{FDC} clearly differed for some strategies as a function of the objective function. All sampling strategies with high runoff values poorly constrained model parameters for FDC simulations when calibrated based on R_{eff}. Using the objective function R_{eff_logQ} for model calibration strongly improved R^*_{FDC} for strategies combining maximum runoff with minimum runoff or with runoff samples at a fixed time interval ($I_{Max\_Min\_Dom}$, $C_{Max\_Rec\_Dom}$, S_{Max_Min} and $C_{Max\_Min\_Wetness}$). Sampling strategies covering low and mean flows (S_{Min}, S_{Mean}, S_{DOM} and $I_{Quantile}$) mostly led to good R^*_{FDC} values with slightly higher model performance for calibrations based on R_{eff_logQ} (R^*_{FDC} from 0.78 to 0.92). Model calibration on R_{eff_logQ} guided parameter selection in a way that some sampling strategies provided informative runoff samples for both hydrograph and FDC, whereas the value of sampling strategies was restricted to either of these simulation aims for calibrations with R_{eff} (Fig. 5a and b).

Model performance generally varied greatly between calibration periods for all strategies and catchments (Fig. 6a and b; standard deviation shown on y-axis). However, it was not possible to establish any relation between hydroclimatic conditions (e.g. yearly or seasonal precipitation, runoff or snowfall) or variations in runoff measurement magnitudes and model performance of the calibrated model. The differences in yearly model performance were smaller for model calibrations with informative sampling strategies, which was indicated by the negative correlation between the median model performance and the standard deviation of the model performance for calibrations based on R_{eff_logQ} (Fig. 6a and b). Also, the relative value of sampling strategies for the simulation of the hydrograph or the FDC was consistent over the fourteen calibration periods (Fig. 7).

We were further interested in how sampling strategies constrained the different model parameters during calibration (Fig. 8). Parameters of the snow routine had mostly large normalized parameter ranges for all sampling strategies indicating that model simulations were often not sensitive to the parameter value. This was different for the five catchments with the highest percentage of precipitation falling as snow, where TT, CFMAX and SFCF were clearly better constrained with normalized ranges as low as 0.42, 0.25, and 0.65. Parameters influencing the water balance (soil routine and PERC of groundwater routine) were better constrained by strategies that sample low and mean flow. However, hydrograph related parameters (UZL, K0 and MAXBAS in

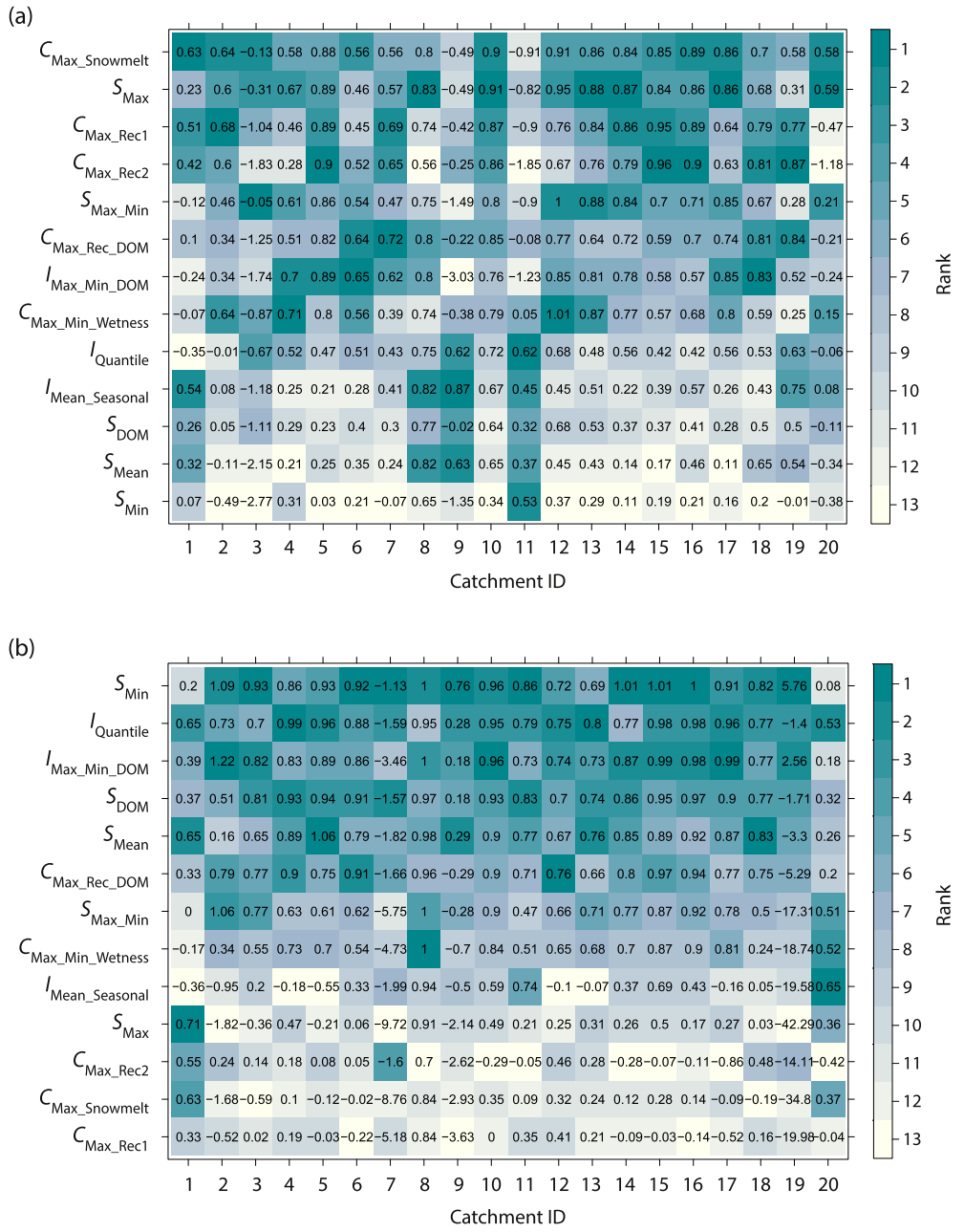


Fig. 4. Normalized model performance as validated for a) the hydrograph (R^*_{eff}) and b) the FDC (R^*_{FDC}) for model calibrations with the sampling strategies using R_{eff_logQ} as objective function. The normalized performance values correspond to the median ensemble mean of all calibration years. Sampling strategies were ranked according to their model performance. Sampling strategies on the y-axis are ordered by their mean rank over all catchments. Colours indicate the rank of a sampling strategy for a particular catchment. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

the groundwater and routing routine) were generally more similar if the model was calibrated with sampling strategies containing maximum runoff.

4. Discussion

The modelling results indicate that a limited number of strategically selected runoff samples is informative for hydrograph and FDC simulations in almost ungauged catchments. Different combinations of runoff samples had a different information value for simulating the hydrograph and the FDC. Possible factors contributing to this difference could be the runoff distribution resulting from a particular sampling strategy (boxplots in Fig. 2) and the model parameters most sensitive at the point in time a runoff sample was provided for calibration. Model parameters of the groundwater and the routing routine that define the timing and the shape of the hydrograph had the least uncertainty when the model was calibrated with runoff samples of high flows and recessions. The benefit of maximum runoff and event data for model calibration was also reported by Seibert and Beven (2009) and Seibert and McDonnell (2013). Our results also confirm the conclusion of several studies (Yapo et al., 1996; Vrugt et al., 2006; Kim and Kaluarachchi, 2009; Melsen et al., 2014; Correa et al., 2016) that rather average and dry runoff periods, represented by samples of mean and minima flows, are less informative for hydrograph prediction than wet periods. For FDC simulations it is crucial to accurately model runoff magnitudes, whereas the exact shape of the hydrograph is less important. Therefore, sampling strategies

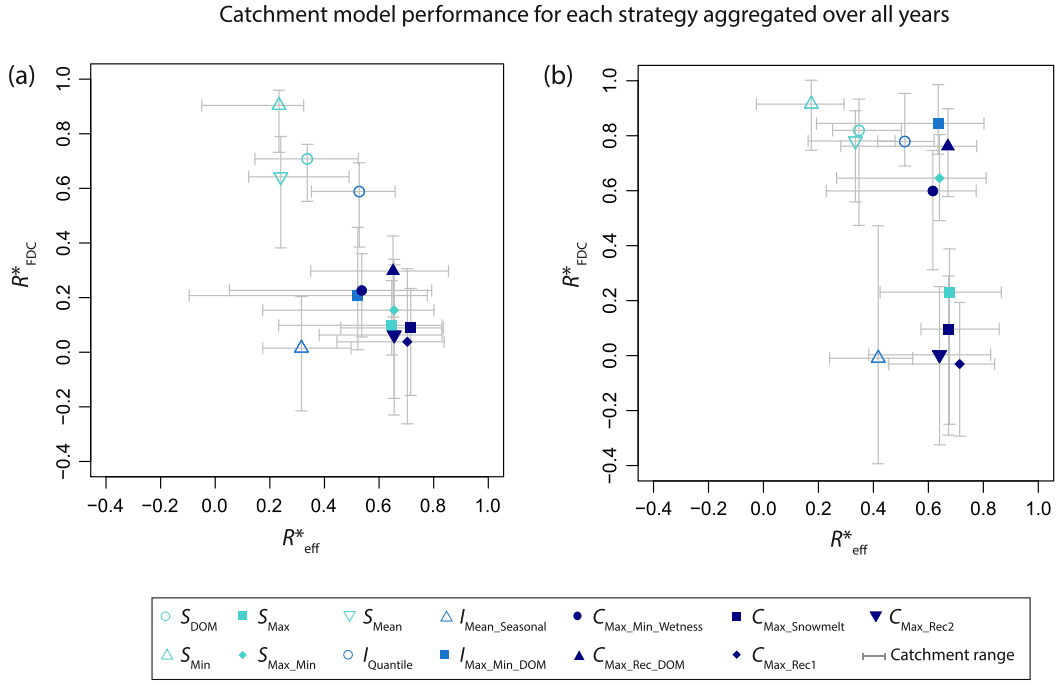


Fig. 5. Normalized model performance as validated for the hydrograph (R^*_{eff}) and the FDC (R^*_{FDC}) for model calibrations with the sampling strategies using (a) R_{eff} and (b) R_{eff_logQ} as objective functions. Each symbol represents the median model performance for a particular strategy over all catchments. It was calculated on the basis of the median ensemble mean of all calibration years. Error bars indicate the 0.25–0.75 quantile model performance of all catchments for the respective strategy. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

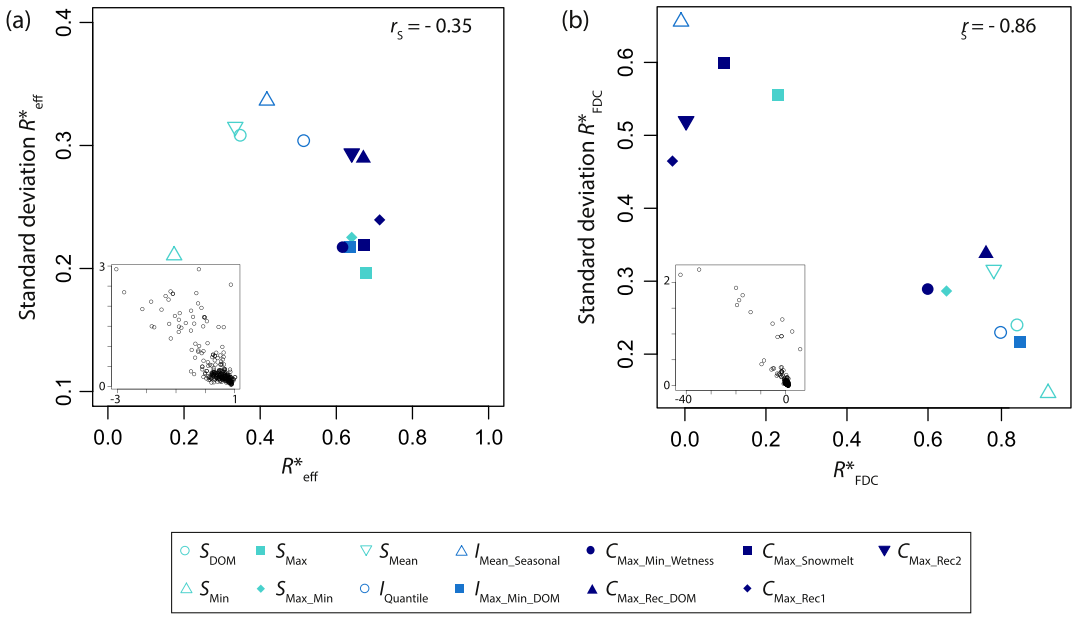


Fig. 6. Comparison of the normalized model performance and the standard deviation of the normalized model performance as validated for a) the hydrograph (R^*_{eff}) and b) the FDC (R^*_{FDC}) for model calibrations with the sampling strategies using R_{eff_logQ} as objective function. Each coloured symbol represents the median model performance and the median standard deviation of the model performance for a particular strategy over all catchments. The median and the standard deviation were calculated on the basis of the ensemble mean of all calibration years. r_S corresponds to the Spearman's rank correlation coefficient between the median R^*_{eff} and the standard deviation of R^*_{eff}. The inset plot makes the same comparison, but indicating the values for each catchment separately. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

resulting in a comparable runoff distribution as a continuous long-term runoff time series were most valuable for simulating the FDC. These strategies, e.g. S_{DOM}, S_{Mean} or $I_{Quantile}$, were most effective in constraining parameters with strong impact on the water balance (soil routine and percolation parameters). None of the sampling strategies noticeably reduced the high uncertainty of snow related model parameters, probably because many study catchments had no or little snowfall.

It is interesting that strategies combining samples of maximum, minimum and recession flow could become informative for the

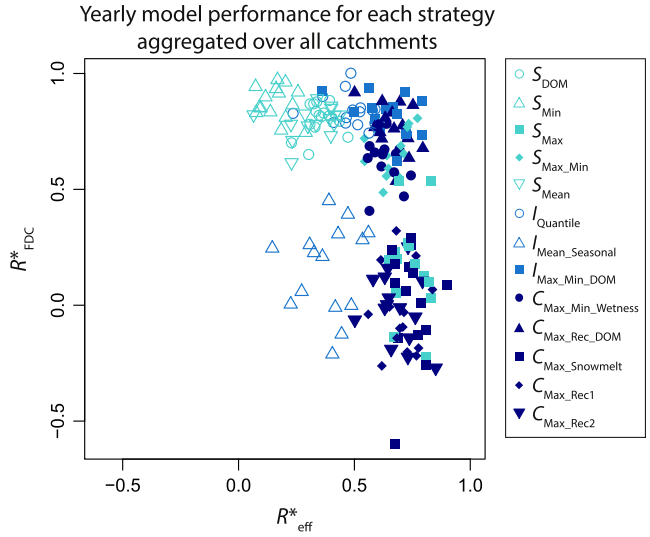


Fig. 7. Normalized model performance as validated for the hydrograph (R^*_{eff}) and the FDC (R^*_{FDC}) for model calibrations with the sampling strategies using R_{eff_logQ} as objective function. Each symbol represents the median model performance for a particular strategy over all catchments for one calibration year. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

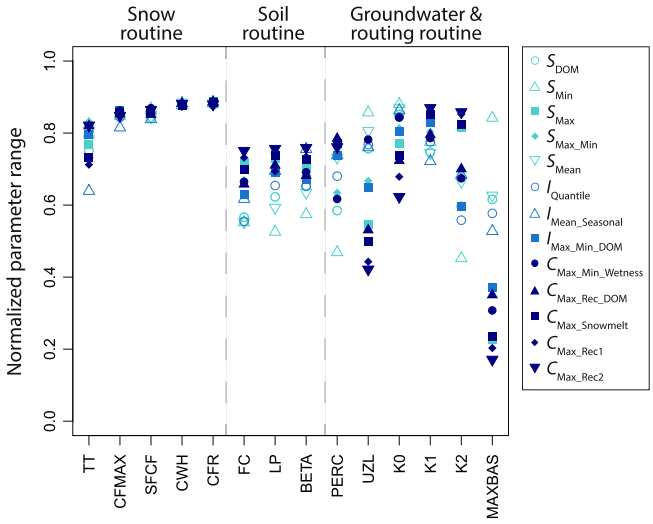


Fig. 8. Normalized model parameter ranges resulting from model calibrations with the sampling strategies using R_{eff_logQ} as objective function. Parameter ranges (0.05–0.95 quantile) after calibration were normalized by their allowed range before calibration. The symbols represent the median normalized parameter range of all catchments related to a particular strategy. This range was calculated on the basis of the median normalized parameter range of all calibration years. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

prediction of the FDC when HBV was calibrated with R_{eff_logQ} instead of R_{eff}. This considerable change could be explained by the distinct focus of the two objective functions during calibration. R_{eff_logQ} emphasises low and mean flow giving more weight to the accurate simulation of a range of magnitudes, while the timing of peak flows is of minor importance. This result demonstrates the importance of carefully choosing the objective function used to optimize model simulations.

The ranking of sampling strategies according to their related model performance (Fig. 4a and b) was clearly less consistent between the twenty catchments for the hydrograph than for the FDC. We tested various catchment characteristics to explain these ranking differences, but no variable was found that could clearly explain the results. Similarly, it was not possible to establish consistently strong relationships between catchment characteristics and the yearly model performance. The sample of twenty catchments might have been too small to find strong relationships between catchment characteristics and model performance as observed by Perrin et al. (2007) in a comparable modelling study framework.

In this study we decided to analyse the modelling results in relation to benchmarks instead of focusing on absolute model performance values. As suggested by Girons Lopez and Seibert (2016), we related model performance based on a limited number of runoff measurements to model calibrations of a well and a non-informed situation. The concept of benchmarks is especially beneficial when predicting runoff for almost ungauged catchments, where the value of taking a few runoff measurements compared to investing efforts in long-term gauging stations is of interest. Absolute model performance becomes more important for practical applications as efficiencies are too low for a reasonable runoff simulation. At this point it is also important to note that low normalized performance does not imply a poor model calibration. For example, the catchments C3, C9 and C11 had many negative normalized performance values due to high Monte Carlo efficiencies. However hydrographs of these catchments were all well simulated in absolute terms. We would also like to stress that the interpretation of the results was not affected by the use of benchmarked performances, because the normalization of model performance did not change the hierarchy of the thirteen sampling strategies within a catchment.

The proposed sampling strategy approach was implemented assuming that one can take a runoff measurement exactly at a certain point in time, such as at the monthly maximum runoff. This is not possible in practice as the runoff is not known at the beginning of a month or a year. The results in our study give an indication of what could be achieved at best and the question is how much the results might have been affected when the runoff was observed at slightly different points in time. Our modelling results suggested that there is some flexibility in taking runoff samples, because none of the tested sampling strategies proved to be superior for model calibration. In the case of hydrograph prediction it was most important to sample high flows preferably in combination with recession data. The most informative sampling strategies for simulating the FDC are not very time sensitive and it was more essential to sample a representative runoff distribution of the particular catchment.

5. Conclusion

This study evaluated the information value of a small number of runoff measurements for calibrating a runoff model for almost ungauged catchments. Our calibration approach has some interesting implications for the prediction of runoff in almost ungauged catchments. It shows the potential of calibrating a runoff model with as few as twelve strategically sampled runoff measurements. Since the exact timing of taking runoff samples was not a major constraint for model calibration, taking samples could be a realistic and efficient alternative to installing a long-term gauging station. Additionally, we applied a runoff model that only requires daily temperature, precipitation and monthly potential evaporation as input, which are variables often available in many regions around the world. The proposed calibration approach could therefore be especially valuable for water management decisions and the mitigation of natural hazards in data scarce regions. However, in case of remote catchments, it might not be time and cost effective to take twelve runoff samples distributed over a hydrological year.

Different strategies for sampling runoff at higher time resolutions within the duration of a short field campaign could be tested to evaluate the value of data for these catchments. Furthermore, our results are limited to humid catchments with little precipitation seasonality and dominated by rain or snow processes. Further investigations are required to evaluate the value of individual runoff measurements, for e.g., arid and glaciated catchments or catchments with a marked precipitation seasonality.

Acknowledgements

The authors thank all the people taking the time to discuss this study and supporting us with new ideas and comments for sampling strategies. We also thank the University of Zurich for funding this work. The two anonymous reviewers are acknowledged for their valuable comments on an earlier draft of this manuscript. Hydrometerological data and catchment shapefiles were made available from Newman et al. (2015). SRTM elevation data was used from Jarvis et al. (2008).

References

Beven, K.J., 2012. Rainfall-Runoff Modelling – The Primer. Wiley and Sons, Chichester.

Bergström, S., 1976. Development and Application of a Conceptual Runoff Model for Scandinavian Catchments. SMHI, Norrköping, Sweden, No. RHO 7, pp 134.

Correa, A., Windhorst, D., Crespo, P., Célleri, R., Feyen, J., Breuer, L., 2016. Continuous versus event-based sampling: how many samples are required for deriving general hydrological understanding on Ecuador's páramo region? Hydrol. Process. 30, 4059–4073. https://doi.org/10.1002/hyp.10975.

Coopersmith, E.J., Minsker, B.S., Sivapalan, M., 2014. Patterns of regional hydroclimatic shifts: an analysis of changing hydrologic regimes. Water Resour. Res. 50, 1960–1983. https://doi.org/10.1002/2012WR013320.

Drogue, G.P., Plasse, J., 2014. How can a few streamflow measurements help to predict daily hydrographs at almost ungauged sites? Hydrol. Sci. J. 59, 2126–2142. https://doi.org/10.1080/02626667.2013.865031.

ESRI and U.S. Geological Survey, 2011. Shaded relief, medium resolution, USA. Available at: ArcGIS online maps and data, last access: March 2017.

Girons Lopez, M., Seibert, J., 2016. Influence of hydro-meteorological data spatial aggregation on streamflow modelling. J. Hydrol. 541, 1212–1220. https://doi.org/10.1016/j.jhydrol.2016.08.026.

Freer, J.E., McMillan, H., McDonnell, J.J., Beven, K.J., 2004. Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. J. Hydrol. 291, 254–277. https://doi.org/10.1016/j.jhydrol.2003.12.037.

Harlin, J., 1991. Development of a process oriented calibration scheme for the HBV hydrological model. Nordic Hydrol. 22, 15–36.

Hrachowitz, M., Savenije, H.H.G., Blöschl, G., McDonnell, J.J., Sivapalan, M., Pomeroy, J.W., Arheimer, B., Blume, T., Clark, M.P., Ehret, U., Fenicia, F., Freer, J.E., Gelfan, A., Gupta, H.V., Hughes, D.A., Hut, R.W., Montanari, A., Pande, S., Tetzlaff, D., Troch, P.A., Uhlenbrook, S., Wagener, T., Winsemius, H.C., Woods, R.A., Zehe, E., Cudennec, C., 2013. A decade of predictions in ungauged basins (PUB): a review. Hydrol. Sci. J. 58, 1198–1255. https://doi.org/10.1080/02626667.2013.803183.

Hughes, D.A., Gush, M., Tanner, J., Dye, P., 2014. Using targeted short-term field investigations to calibrate and evaluate the structure of a hydrological model. Hydrol. Process. 28, 2794–2809. https://doi.org/10.1002/hyp.9807.

Jarvis, A., Reuter, H.I., Nelson, A., Guevara, E., 2008. Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m. Available at: http://srtm.csi.cgiar.org, last access: January 2016.

Juston, J., Seibert, J., Johansson, P.O., 2009. Temporal sampling strategies and uncertainty in calibrating a conceptual hydrological model for a small boreal catchment. Hydrol. Process. 23, 3093–3109. https://doi.org/10.1002/hyp.7421.

Kim, U., Kaluarachchi, J.J., 2009. Hydrologic model calibration using discontinuous data: an example from the upper Blue Nile River Basin of Ethiopia. Hydrol. Process. 23, 3705–3717. https://doi.org/10.1002/hyp.7465.

Konz, M., Seibert, J., 2010. On the value of glacier mass balances for hydrological model calibration. J. Hydrol. 385, 238–246. https://doi.org/10.1016/j.jhydrol.2010.02.025.

Lehner, B., Döll, P., 2004. Development and validation of a global database of lakes, reservoirs and wetlands. J. Hydrol. 296, 1–22. https://doi.org/10.1016/j.jhydrol.2004.03.028.

Lindström, G., Johansson, B., Persson, M., Gardelin, M., Bergström, S., 1997. Development and test of the distributed HBV-96 hydrological model. J. Hydrol. 201, 272–288. https://doi.org/10.1016/S0022-1694(97)00041-3.

McIntyre, N.R., Wheater, H.S., 2004. Calibration of an in-river phosphorus model: prior evaluation of data needs and model uncertainty. J. Hydrol. 290, 100–116. https://doi.org/10.1016/j.jhydrol.2003.12.003.

Melsen, L.A., Teuling, A.J., Berkum, S.W., Torfs, P.J.J.F., Uijlenhoet, R., 2014. Catchments as simple dynamical systems: a case study on methods and data requirements for parameter identification. Water Resour. Res. 50, 5577–5596. https://doi.org/10.1002/2013WR014720.

Merz, R., Parajka, J., Blöschl, G., 2009. Scale effects in conceptual hydrological modeling. Water Resour. Res. 45, W09405. https://doi.org/10.1029/2009WR007872.

Newman, A.J., Clark, M.P., Sampson, K., Wood, A., Hay, L.E., Bock, A., Viger, R.J., Blodgett, D., Brekke, L., Arnold, J.R., Hopson, T., Duan, Q., 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrol. Earth Syst. Sci. 19, 209–223. https://doi.org/10.5194/hess-19-209-2015.

Parajka, J., Viglione, A., Rogger, M., Salinas, J.L., Sivapalan, M., Blöschl, G., 2013. Comparative assessment of predictions in ungauged basins-Part 1: Runoff-hydrograph studies. Hydrol. Earth Syst. Sci. 17, 783–1795. https://doi.org/10.5194/hess-17-1783-2013.

Perrin, C., Oudin, L., Andreassian, V., Rojas-Serna, C., Michel, C., Mathevet, T., 2007. Impact of limited streamflow data on the efficiency and the parameters of rainfall-runoff models. Hydrol. Sci. J. 52, 131–151. https://doi.org/10.1623/hysj.52.1.131.

Rojas-Serna, C., Lebecherel, L., Perrin, C., Andreassian, V., Oudin, L., 2016. How should a rainfall-runoff model be parameterized in an almost ungauged catchment? A methodology tested on 609 catchments. Water Resour. Res. 52, 4765–4784. https://doi.org/10.1002/2015WR018549.

Rojas-Serna, C., Michel, C., Perrin, C., Andreassian, V., Hall, A., Chahinian, N., Schaake, J., 2006. Ungauged catchments: How to make the most of a few streamflow measurements? Large sample basin experiments for hydrological model parameterization: results of the model parameter experiment – MOPEX. IAHS Publ. 307, 230–236.

Seibert, J., Beven, K.J., 2009. Gauging the ungauged basin: How many discharge measurements are needed? Hydrol. Earth Syst. Sci. 13, 883–892. https://doi.org/10.5194/hess-13-883-2009.

Seibert, J., McDonnell, J.J., 2013. Gauging the ungauged basin: relative value of soft and hard data. J. Hydrol. Eng. 20, A4014004. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000861.

Seibert, J., Vis, M.J.P., 2012. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. Hydrol. Earth Syst. Sci. 16, 3315–3325. https://doi.org/10.5194/hess-16-3315-2012.

Singh, S.K., Bárdossy, A., 2012. Calibration of hydrological models on hydrologically unusual events. Adv. Water Resour. 38, 81–91. https://doi.org/10.1016/j.advwatres.2011.12.006.

Sivapalan, M., Takeuchi, K., Franks, S.W., Gupta, V.K., Karambiri, H., Lakshim, V., Liang, X., McDonnell, J.J., Mendiondo, E.M., Connell, O., Oki, T., Pomeroy, J.W., Schertzer, D., Uhlenbrook, S., Zehe, E., 2003. IAHS decade on Predictions in Ungauged Basins (PUB), 2003–2012: shaping an exciting future for the hydrological sciences. Hydrol. Sci. J. 48, 857–880. https://doi.org/10.1623/hysj.48.6.857.51421.

Sun, W., Wang, Y., Wang, G., Cui, X., Yu, J., Zuo, D., Xu, Z., 2017. Physically based distributed hydrological model calibration based on a short period of streamflow data: case studies in four Chinese basins. Hydrol. Earth Syst. Sci. 21, 251–265. https://doi.org/10.5194/hess-21-251-2017.

Uhlenbrook, S., Sieber, A., 2005. On the value of experimental data to reduce the prediction uncertainty of a process-oriented catchment model. Environ. Modell. Softw. 20, 19–32. https://doi.org/10.1016/j.envsoft.2003.12.006.

U.S. Geological Survey, 2014. EflowStats R-package. Available at: https://github.com/USGS-R/EflowStats, last access: July 2016

Viviroli, D., Seibert, J., 2015. Can a regionalized model parameterisation be improved with a limited number of runoff measurements? J. Hydrol. 529, 49–61. https://doi.org/10.1016/j.jhydrol.2015.07.009.

Vrugt, J.A., Gupta, H.V., Dekker, S.C., Sorooshian, S., Wagener, T., Bouten, W., 2006. Application of stochastic parameter optimization to the Sacramento soil moisture accounting model. J. Hydrol. 325, 288–307. https://doi.org/10.1016/j.jhydrol.2005.10.041.

Westerberg, I.K., Guerrero, J.L., Younger, P.M., Beven, K.J., Seibert, J., Halldin, S., Freer, J.E., Xu, C.Y., 2011. Calibration of hydrological models using flow-duration curves. Hydrol. Earth Syst. Sci. 15, 2205–2227. https://doi.org/10.5194/hess-15-2205-2011.

Xia, Y., Yang, Z.L., Jackson, C., Stoffa, P.L., Sen, M.K., 2004. Impacts of data length on optimal parameter and uncertainty estimation of a land surface model. J. Geophys. Res. Atmos. 109, D07101. https://doi.org/10.1029/2003JD004419.

Yapo, P.O., Gupta, H.V., Sorooshian, S., 1996. Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. J. Hydrol. 181, 23–48. https://doi.org/10.1016/0022-1694(95)02918-4.

Paper V

Resubmitted after minor revisions to *Water Resources Research*

Value of a Limited Number of Discharge Observations for Improving Regionalisation: A Large-Sample Study across the United States

S. Pool[1], D. Viviroli[1], and J. Seibert[1, 2]

[1]University of Zurich, Department of Geography, Zurich, Switzerland

[2] Swedish University of Agricultural Sciences, Department of Aquatic Sciences and Assessment, Uppsala, Sweden

Corresponding author: Sandra Pool (sandra.pool@geo.uzh.ch)

Abstract: Even in regions considered as densely monitored, most catchments are actually ungauged. Prediction of discharge in such ungauged catchments commonly relies on parameter regionalisation. While ungauged catchments lack continuous discharge time series, a limited number of observations could still be collected within short field campaigns. Here, we analyse the value of such observations for improving parameter regionalisation in otherwise ungauged catchments. More specifically, we propose an ensemble modelling approach, where discharge predictions from regionalisation with multiple donor catchments are weighted based on the fit between predicted and observed discharge on the dates of the available observations. It was assumed that a total of 3 to 24 observations from a single hydrological year were available as an additional source of information for regionalisation. This informed regionalisation approach was tested with discharge observations from 10 different hydrological years in a leave-one-out cross validation scheme on 579 catchments in the United States using the HBV runoff model. Discharge observations helped to improve the regionalisation in up to 94 % of the study catchments in 8 out of 10 discharge sampling years. Sampling years characterized by exceptionally high peak discharge, or high annual or winter precipitation were less informative for regionalisation. In the least informative years, model efficiency increased with an increasing number of observations. In contrast, in the most informative sampling year, 3 discharge observations provided as much information for regionalisation as 24 discharge observations. Overall, discharge observations were most effective in informing regionalisation in arid catchments, snow dominated catchments and winter-precipitation dominated catchments.

Keywords: Ungauged basin, regionalization, spatial proximity, attribute similarity, value of data

1 Introduction

Continuous discharge time series are fundamental for many water management decisions in a river basin. However, even in regions considered as densely monitored, a considerable fraction of catchments are ungauged or poorly gauged, i.e. have no or only limited discharge data. Estimates of continuous discharge time series in such catchments are often based on runoff models, which contain of a number of tuneable parameters that are typically derived from calibration against observed discharge. Determining these model parameters for data scarce catchments is one of the major challenges in hydrology (Hrachowitz et al., 2013).

In the absence of any discharge data, model parameters can be estimated using regionalisation methods, whereby hydrologic information is transferred from gauged to ungauged locations (Blöschl & Sivapalan, 1995). Regionalisation is a long-standing research question in hydrology and has received special attention due to the PUB (Prediction in Ungauged Basins) initiative (Sivapalan et al., 2003). There are a great number of regionalisation approaches that have been proposed (for reviews see e.g. He et al., 2011; Parajka et al., 2013; Razavi & Coulibaly, 2013). Although the most suitable regionalisation approach is likely site-specific (He et al., 2011; Razavi et al., 2013), it has been argued that approaches that transfer parameters as a set rather than individually are favourable since they account for parameter dependence (Bárdossy, 2007; Buytaert & Beven, 2009; Kokkonen et al., 2003; McIntyre et al., 2005). Moreover, regionalisation performance is higher when averaging discharge simulations from parameter sets of multiple donor catchments as opposed to the selection of one single donor (Arsenault & Brissette, 2014; Oudin et al., 2008; Yang et al., 2018; Zhang & Chiew, 2009). Similarly, the combination of multiple regionalisation methods can outperform predictions based on a single approach (Oudin et al., 2008; Viviroli et al., 2009; Yang et al., 2018; Zhang & Chiew, 2009).

Spatial proximity and attribute similarity are among the most commonly applied regionalisation approaches that use entire parameter sets from one or multiple donor catchment(s). Spatial proximity is based on Tobler's first law of Geography that "everything is related to everything else, but near things are more related than distant things" (Tobler, 1970, p.236). In the context of hydrological modelling it can be assumed that climate and catchment attributes vary smoothly in space (Parajka et al., 2013). Therefore, the distance between catchment outlets (Parajka et al., 2013), centroids (Arsenault & Brissette, 2014; Oudin et al., 2008; Samuel et al., 2011) or a combination thereof (Lebecherel et al., 2016) can be used to select hydrologically similar donor catchments. The efficiency of the spatial proximity approach obviously depends on the density of the streamflow gauging network (Lebecherel et al., 2016; Samuel et al., 2011). Attribute similarity-based regionalisation approaches presume that the degree of similarity between catchments can be expressed by a multitude of catchment attributes that are linked to a catchment's runoff

response (Burn & Boorman, 1993). Commonly used attributes quantify topographical characteristics, land cover, climatic conditions, or soil characteristics and geology (Arsenault & Brissette, 2014; Merz & Blöschl, 2004; Oudin et al., 2008; Viviroli et al., 2009; Zhang & Chiew, 2009).

In some cases a catchment of interest lacks long continuous discharge time series, but a small number of discharge observations could be collected within a limited time period. Several studies (Melsen et al., 2014; Seibert & Beven, 2009; Seibert & McDonnell, 2015; Singh & Bárdossy, 2012) have shown the value of a limited number of discharge observations for model calibration. Observations during wet periods (Melsen et al., 2014; Vrugt et al., 2006; Yapo et al., 1996) or at an event peak and the subsequent recession limb (Pool et al., 2017; Seibert & Beven, 2009; Seibert & McDonnell, 2015) are particularly informative for parameter estimation. Furthermore, a limited number of observations was shown to be most informative if it represents the dominant hydrological processes and covers a range of hydrological conditions (Harlin, 1991; Singh & Bárdossy, 2012; Vrugt et al., 2006).

A few available discharge observations could also be used in combination with parameter regionalisation. For example, Viviroli and Seibert (2015) weighted parameter sets from each donor catchment based on their ability to reproduce discharge observations taken during average flow conditions. They tested the proposed approach on 49 catchments in Switzerland and report that a few observations can improve discharge predictions, especially for snow and glacier dominated catchments. In a comparable parameter weighting approach, Rojas-Serna et al. (2016) analysed the value of a varying number of randomly selected discharge observations for regionalisation. Results were based on 609 catchments in France and indicate that 5 discharge observations can already be informative for regionalisation and that 10 observations increased model efficiency by up to 50 %.

In this study, we assumed that a limited number of discharge observations were available for an ungauged catchment. These observations were then used to inform parameter estimation together with regionalisation based on spatial proximity or attribute similarity. The proposed approach was tested with the HBV runoff model (Bergström, 1976; Lindström et al., 1997) and the CAMELS data set (Addor et al., 2017; Newman et al., 2015) using a leave-one-out cross validation, i.e. treating each catchment ungauged in turns. Our specific research questions were as follows:

1) Can a limited number of discharge observations be used to improve regionalisation?
2) Does the value of a limited number of discharge observations vary between different types of catchments?
3) How much does the value of discharge observations vary between different sampling years?
4) How much does the value of discharge observations change with the number of observations?

2 Study area and data

This study was based on 579 catchments from across the contiguous United States (Fig. 1). The catchments cover a large range of hydroclimatic and landscape characteristics (Table 1). The study catchments are a subset (see Chpt. 3.2) of the publicly available CAMELS data set (version 1.0; Addor et al., 2017; Newman et al., 2015). CAMELS consists of over 600 catchments in the Unites States with minimum human disturbance. The data set provides 20 year long time series with daily discharge and meteorological data for each catchment. Moreover, it includes catchment boundaries along with a list of 80 catchment descriptors, such as location and topography attributes, climate indices, soil characteristics, and vegetation characteristics. As an additional catchment characteristic, we extracted the percentage of catchment area classified as wetlands from the global data set of Lehner & Döll (2004). Using the meteorological data provided by CAMELS, we furthermore calculated the monthly potential evaporation based on the Priestley-Taylor equation (Priestley & Taylor, 1972).

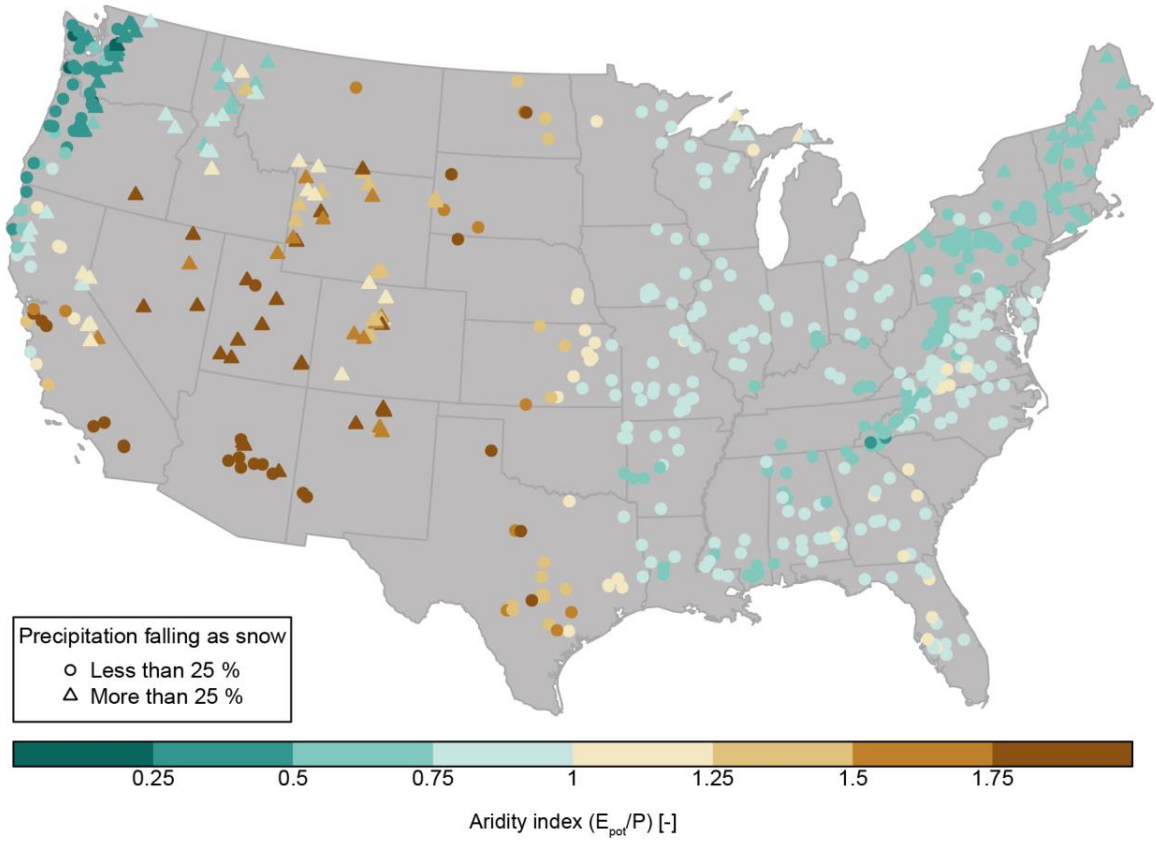


Figure 1. Locations of the 579 study catchments. Colours indicate the aridity index and the marker shape denotes the percentage of precipitation falling as snow.

Table 1. Statistics of Catchment Attributes Used in this Study.

Catchment attribute	5th quantile	Median	95th quantile
Area [km^2]	22	301	2432
Aridity index[a] [-]	0.33	0.83	1.94
Precipitation seasonality[b] [-]	-1.13	0.06	0.65
Precipitation falling as snow [%]	0	9	69
Forested area [%]	2	86	100
Wetland area[c] [%]	0	0	96
Clay content in soils [%]	6	19	36

Note: [a]Aridity index equals the ratio of sum of potential evaporation and sum of precipitation; [b]Precipitation seasonality is negative for catchments with winter precipitation, zero for catchments without precipitation seasonality, and positive for catchments with summer precipitation (for the calculation see Addor et al., 2017); [c]Percentage of catchment area classified as wetland was extracted from the global data set of Lehner & Döll (2004).

3 Model structure and model calibration

3.1 HBV model

Continuous daily discharge time series were simulated with the HBV model (Bergström, 1976; Lindström et al., 1997) in the version HBV-light (Seibert & Vis, 2012). HBV is a bucket-type runoff model that consists of four routines with a conceptual representation of snow pack dynamics, soil moisture variation, runoff response and discharge routing. The model is forced by daily temperature and precipitation data as well as monthly potential evaporation data. In the snow routine, precipitation is assumed to fall as snow and accumulates as soon as temperatures drop below a threshold value. Snowmelt and refreezing of liquid snow water content are both estimated based on the degree-day method. Snowmelt, rainfall, and potential evaporation are inputs to the soil routine, where actual evaporation and recharge to the groundwater are determined as a function of the simulated soil moisture storage. The groundwater routine consists of a shallow and a deep storage that contributes to the peak, intermediate and baseflow components of the hydrograph. Finally, the three discharge components are summed and transformed into the hydrograph at the catchment outlet by a triangular weighting function.

In this study, HBV was used in a semi-distributed form by disaggregating each catchment into elevation bands of 200 m using SRTM elevation data (Jarvis et al., 2008). Hydrological processes in the snow and the soil routine were calculated separately for each elevation band, whereas groundwater was represented as a single storage over the entire catchment. The area-weighted mean precipitation and temperature input

from the CAMELS data set were interpolated across elevation bands using a constant lapse rate of 10 % per 100 m and 0.6°C per 100 m, respectively. Potential evaporation was assumed to be constant within each catchment.

3.2 Model calibration

The HBV model was calibrated for each study catchment using meteorological input and continuous daily discharge time series from 1 October 1989 to 30 September 1999. A warm-up period of 2 ¾ years preceded the calibration period to ensure suitable initial values for the state variables. Model parameters were optimized within predefined parameter ranges using a genetic algorithm (Seibert, 2000) and a modified variant of the Kling-Gupta efficiency (Gupta et al., 2009) towards a non-parametric metric as objective function (R_{NP}; Pool et al., *in press*). Similar to the King-Gupta efficiency, R_{NP} (Eq. 1) consists of the three error terms representing discharge volume (β; Eq. 2), variability (α_{NP}; Eq. 3), and dynamics (r_S; Eq. 4). In the equations, β is the bias between observed (*obs*) and simulated (*sim*) mean discharge μ, α_{NP} is the absolute error between the observed and simulated normalized flow-duration curve (where $I(k)$ and $J(k)$ are the time steps when the kth largest flow (Q) occurs within the simulated and observed time series, respectively), and r_S corresponds to the Spearman rank correlation between the ranks of the observed (R_{obs}) and simulated (R_{sim}) discharge time series at time step t.

$$R_{NP} = 1 - \sqrt{(\beta - 1)^2 + (\alpha_{NP} - 1)^2 + (r_s - 1)^2} \quad \text{Eq. (1)}$$

$$\beta = \frac{\mu_{sim}}{\mu_{obs}} \quad \text{Eq. (2)}$$

$$\alpha_{NP} = 1 - \frac{1}{2}\sum_{k=1}^{n}\left|\frac{Q_{sim}(I(k))}{n\bar{Q}_{sim}} - \frac{Q_{obs}(J(k))}{n\bar{Q}_{obs}}\right| \quad \text{Eq. (3)}$$

$$r_s = \frac{\sum_{i=1}^{n}(R_{obs}(t) - \bar{R}_{obs})(R_{sim}(t) - \bar{R}_{sim})}{\sqrt{(\sum_{i=1}^{n}(R_{obs}(t) - \bar{R}_{obs})^2)(\sum_{i=1}^{n}(R_{sim}(t) - \bar{R}_{sim})^2)}} \quad \text{Eq. (4)}$$

To account for parameter uncertainty and equifinality (Beven & Freer, 2001), we calibrated the HBV model a 100 times, resulting in 100 optimized parameter sets for each catchment. Catchments for which the model failed to reproduce discharge at an acceptable level were discarded from the further regionalisation as suggested by Arsenault and Brissette (2014) and Bárdossy (2007). For the level of acceptance, we applied a threshold of $R_{NP} > 0.65$, which is comparable to a Nash-Sutcliffe model efficiency (Nash & Sutcliffe, 1970) of $R_{NS} > \sim 0.2$. This selection resulted in the final set of 579 study catchments from the originally more than 600 catchments of the CAMELS data set.

4 Modelling framework

4.1 Regionalisation methods

In this study, regionalisation was based on five donor catchments. These donors were selected by defining homogeneous regions for every single catchment, which is known as the region of influence approach (Burn, 1990). Homogeneous regions consist of similar catchments and were defined as i) regions containing spatially close catchments or ii) regions with catchments having similar attributes. Spatial proximity was defined as the Euclidian distance (Burn, 1990; McIntyre et al., 2005) between the coordinates of catchment centroids, whereas attribute similarity was described using the Euclidian distance in the attribute space (Burn, 1990; McIntyre et al., 2005). The attribute space consisted of seven selected catchment characteristics: catchment area (log-transformed values of area were used), aridity, precipitation seasonality as an indicator for seasonal or perennial precipitation, percentage of precipitation falling as snow, percentage of forested area, percentage of wetland area, and percentage of clay content in soils (Table 1). All attributes were standardized using Eq. 1 (Milligan & Cooper, 1988), where Z is the standardized attribute and X is the original attribute value.

$$Z = \frac{X - X_{min}}{X_{max} - X_{min}} \quad \text{Eq. (5)}$$

Regionalisation was evaluated using a leave-one-out cross validation, where each catchment was treated as ungauged at a time and its discharge was estimated with the information from the donor catchments. Each of the five donors provided its 100 parameter sets from calibration to the ungauged catchment. The total of 500 parameter sets was used to predict discharge in the ungauged catchment during the calibration and the validation period (1 October 1999 to 30 September 2009), leading to 500 hydrographs for the ungauged catchment. The 500 discharge simulations (Q_i) were then aggregated into an ensemble mean hydrograph ($\bar{Q}$, Eq. 6). Discharge at each time step t was derived from equally weighting ($W_i = \frac{1}{N}$) each (i) of the total of N parameter sets.

$$\bar{Q}(t) = \sum_{i=1}^{N} Q_i(t) W_i \quad \text{Eq. (6)}$$

The ensemble mean hydrograph was evaluated in terms of R_{NP}. The described regionalisation approach based on attribute similarity or spatial proximity without any further information will be referred to as *classical regionalisation* in this study.

4.2 Gauging the ungauged catchment

To evaluate the value of individual discharge observations, we assumed that a hydrologist gets the opportunity to take a few discharge measurements within one hydrological year in the previously ungauged catchment. Such a sampling campaign was mimicked by extracting the few daily discharge observations from the observed time series of each catchment. The selection of observations was restricted to a hydrological year, but repeated for each of the 10 calibration years (later on referred to as *sampling years*). A varying number *n* of observations were strategically selected based on our experience from previous studies (Pool et al., 2017; Seibert & McDonnell, 2015). The sampling strategy used to select discharge observations included samples of the annual peak discharge, the subsequent days in the recession of the peak, and observations at a fixed day in different months of the year. Depending on the number (*n*) of measurements, the observations were assumed to have been taken as follows (Fig. 2):

- *n* = 3: 1 peak and 2 days in its recession
- *n* = 6: 1 peak and 2 days in its recession combined with observations at the 15th of 3 months
- *n* = 12: 1 peak and 5 days in its recession combined with observations at the 15th of 6 months
- *n* = 24: 2 peaks and 5 days in their recessions combined with observations at the 15th of each month

The strategically taken discharge observations of the 'ungauged' catchment served to evaluate the 500 hydrograph predictions from regionalisation. The root mean square error (R_{RMSE}) between predicted and observed discharge on the dates of the *n* observations was used to compute a weighted ensemble mean hydrograph (Eq. 6) in the validation time period. The weight W_i of each *i*th parameter set was calculated using Eq. 7, where *N* is the total number of *j* parameter sets ($j = 1,2,\ldots, N$) and $R_{\mathrm{RMSE,max}}$ is the highest R_{RMSE} among all parameter sets. Log-transformed values of R_{RMSE} were used to accentuate the difference between the best and the worst parameter set.

$$W_i = \frac{\ln R_{RMSE,\max} - \ln R_{RMSE,i}}{\sum_{j=1}^{N} \ln R_{RMSE,\max} - \ln R_{RMSE,j}} \qquad \text{Eq. (7)}$$

The above described approach uses the information of a few discharge observations and will therefore be referred to as *informed regionalisation*. As for the classical regionalisation, the ensemble mean hydrograph of the informed regionalisation was evaluated based on R_{NP}. To assess the value of discharge observations for regionalisation, we calculated the difference in efficiency (ΔR_{NP}) between the informed regionalisation (IR) and the classical regionalisation (CR) as follows:

$$\Delta R_{NP} = R_{NP\_IR} - R_{NP\_CR} \qquad \text{Eq. (8)}$$

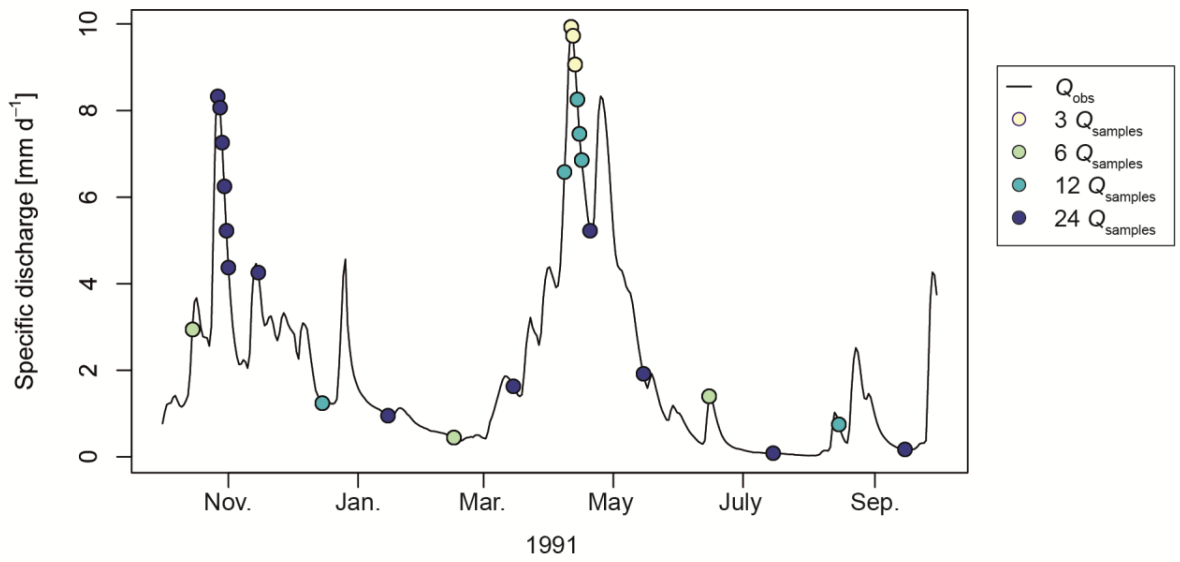


Figure 2. Illustration of a sampling campaign with 3, 6, 12, or 24 discharge observations. The discharge observations ($Q_{samples}$) were extracted from the continuous discharge time series (Q_{obs}) of each catchment, whereby the selection of observations was restricted to a hydrological year. Colours indicate the timing of additional observations that result from increasing the total number of observations taken in a sampling campaign.

4.3 Benchmarks

An upper and a lower benchmark (Seibert et al., 2018) were used as references for the model performance of the classical regionalisation and the informed regionalisation. The upper benchmark was equivalent to the calibration of the model on the continuous 10 year time series. It provides information on how well the model simulates discharge in a particular catchment in a well-informed situation. The lower benchmark indicates the model's ability for simulating discharge in the absence of any discharge information. Simulations for the lower benchmark were run with 10 000 randomly selected parameter sets. The 10 000 parameter sets of the lower benchmark and the 100 parameter sets of the upper benchmark were used to simulate discharge in the validation period. Simulations were again combined into an ensemble mean hydrograph (Eq. 6 with equal weights for all parameter sets) and evaluated using R_{NP}. Additionally, the ensemble mean hydrograph of the upper benchmark served to compute the percentage increase in R_{NP} (ΔR_{NP} divided by the difference between R_{NP} of the upper benchmark and R_{NP} of the classical regionalisation) to have an indication for how close the efficiency of the informed regionalisation is to a well-informed model calibration.

4.4 Variability of model performance in space and time

To investigate regional differences of the value of discharge data, we generated maps of efficiency differences (ΔR_{NP}) and evaluated these differences against catchment attributes. The relation between efficiency differences and catchment attributes is presented for the informed regionalisation with 24 discharge observations and a median sampling year.

The variability of model performance in time was evaluated in two ways. First, we analysed the information content of the 10 sampling years. To this end, sampling years were ranked by the efficiency of the informed regionalisation. The ranks were used to evaluate in how many sampling years discharge observations were informative for the majority of catchments. Second, we compared model efficiencies with the hydrometeorological conditions (e.g. sum of precipitation, peak discharge magnitude) in each sampling year. To enable a comparison between catchments, model efficiencies and hydrometeorological variables were normalized. Model efficiencies were normalized by taking the difference between ΔR_{NP} of a sampling year and the mean ΔR_{NP} of all 10 sampling years, whereas hydrometerological variables were divided by their mean. Correlations between normalized hydrometerological variables and normalized model efficiency difference were quantified by the Spearman rank correlation. Correlations were computed for various subgroups of catchments with different aridity conditions (humid, temperate, and arid), influence of snow-related runoff processes (no snow, more than 15 % of annual precipitation falling as snow, and more than 50 % of annual precipitation falling as snow), and precipitation seasonality (no seasonality, summer precipitation, and winter precipitation).

Lastly, we addressed the question of how many discharge samples are needed to effectively improve classical regionalisation by comparing efficiency differences (ΔR_{NP}) of each sampling year for a different number of discharge samples. Results were evaluated for the average year, as well as for the most and the least informative year, which were determined as described above.

5 Results

5.1 Value of discharge observations for regionalisation

First, we related the value of observations for discharge predictions to model efficiencies of the classical regionalisation and the upper and lower benchmarks (Fig. 3; see supplement for the detailed values). Efficiencies of the classical regionalisation with spatial proximity were about half way between the efficiencies of the upper and lower benchmarks as opposed to efficiencies related to attribute similarity that were closer to the lower benchmark than to the upper benchmark. The use of discharge observations for weighting the 500 parameter sets from the donor catchments improved regionalisation with both attribute

similarity and spatial proximity (Fig. 3), whereby differences in model performance (ΔR_{NP}) between a classical and an informed regionalisation were most pronounced for attribute similarity (Fig. 3b). Three to 24 discharge observations improved model efficiency of classical regionalisation by 24 % to 30 % in case of an attribute-based regionalisation and 22 % to 26 % in case of the spatial proximity-based approach. However, for some catchments the selected discharge observations were disinformative in that the use of information decreased model performance. Such a negative effect occurred mainly for cases with only 3 or 6 observations (for more details see Chpt. 5.3).

Although a comparison of the two classical regionalisation approaches was not the focus of this study, it was interesting to notice that spatial proximity outperformed attribute similarity in 65 % of the catchments. The (frequently) superior efficiency of spatial proximity could also be noted in the fact that regionalisation with attribute similarity had to be informed with 24 discharge observations to reach efficiencies comparable to spatial proximity without any discharge information.

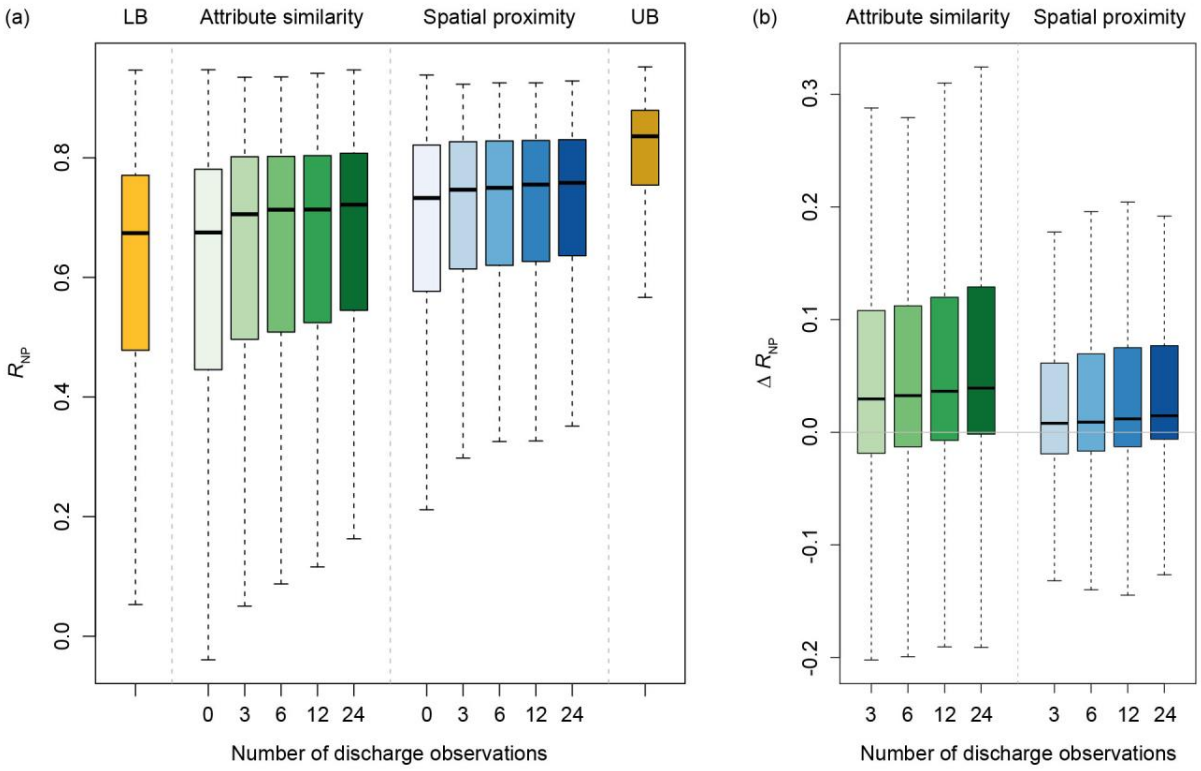


Figure 3. Model efficiency in validation for the 579 study catchments. a) Model efficiency R_{NP} for predictions with the lower (LB) and upper (UB) benchmark (yellow colours), and the classical and the informed regionalisation (efficiency of the median sampling year) with attribute similarity (green colours) and spatial proximity (blue colours). b) Efficiency difference (ΔR_{NP}) between the classical and the informed regionalisation, whereby positive values indicate an increase in prediction efficiency using information of a few discharge observations.

5.2 Mapping the value of discharge observations in space

Mapping the effect of discharge observations on the classical regionalisation allowed to visually separate regions where observations were highly informative from regions where they were less important (Fig. 4). The map suggests that the value of observations varies in space. Discharge observations had no or only limited value in large parts of the central region of the eastern United States, such as the Gulf Coast, the Mississippi Valley and the Great Lakes Region. In contrast, a pronounced positive effect of discharge measurements was observed for the majority of catchments in the Appalachian Mountains and the western United States. The described spatial pattern can also be observed when plotting the effect of discharge observations against catchment attributes (Figs. 5a-g and 6a-g). Discharge observations did in general strongly improve classical regionalisation in arid catchments that are most prominent in the Southwest, and in snow dominated catchments in mountainous regions or northern latitudes. Furthermore, regionalisation with spatial proximity was improved by the information of discharge observations in catchments with a distinct winter precipitation season, which are catchments typically located along the West Coast.

In addition to the variable value of discharge observations as a function of catchment attributes, information of a few observations was also more important when the distance between the ‘ungauged’ catchment and its donors was relatively large (Figs. 5h and 6h).

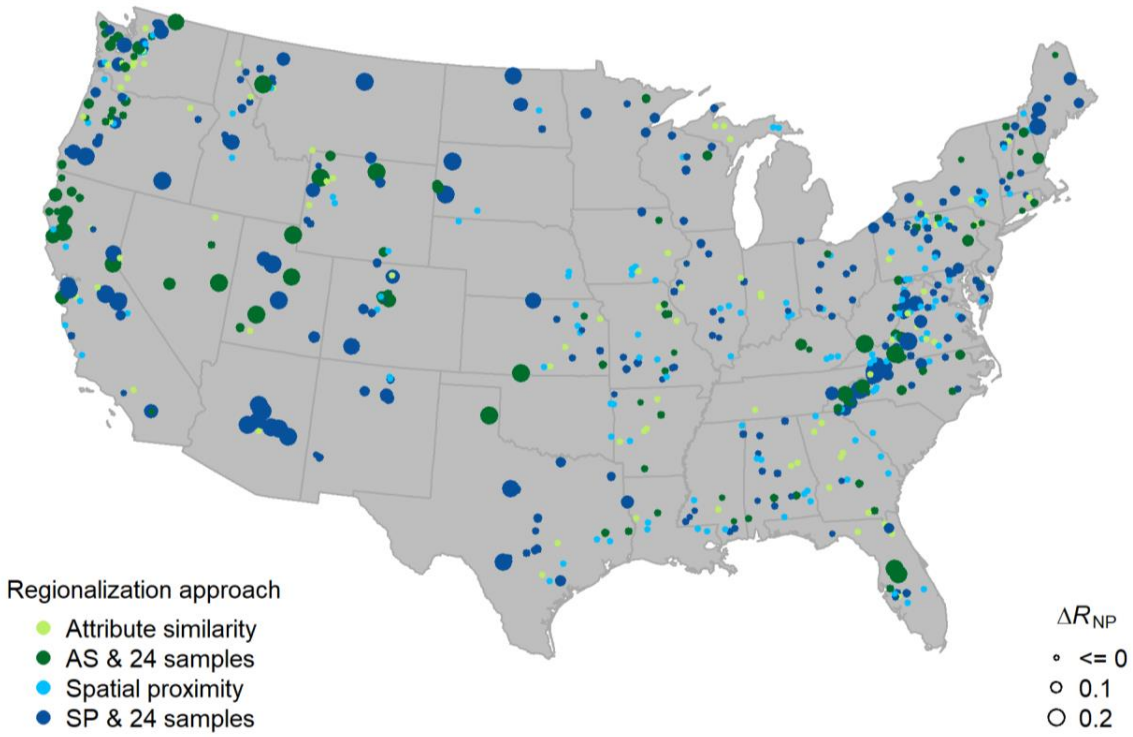


Figure 4. Spatial variability of the difference in validation model efficiency (ΔR_{NP}) between a classical regionalisation and an informed regionalisation with 24 discharge observations (efficiency of the median sampling year). The size of the circles is proportional to ΔR_{NP}, i.e. larger circles indicate a higher value of a few discharge observations for improving classical regionalisation. Green circles denote catchments which were best simulated using regionalisation with attribute similarity (AS), whereas blue circles indicate catchments which were best simulated using regionalisation with spatial proximity (SP).

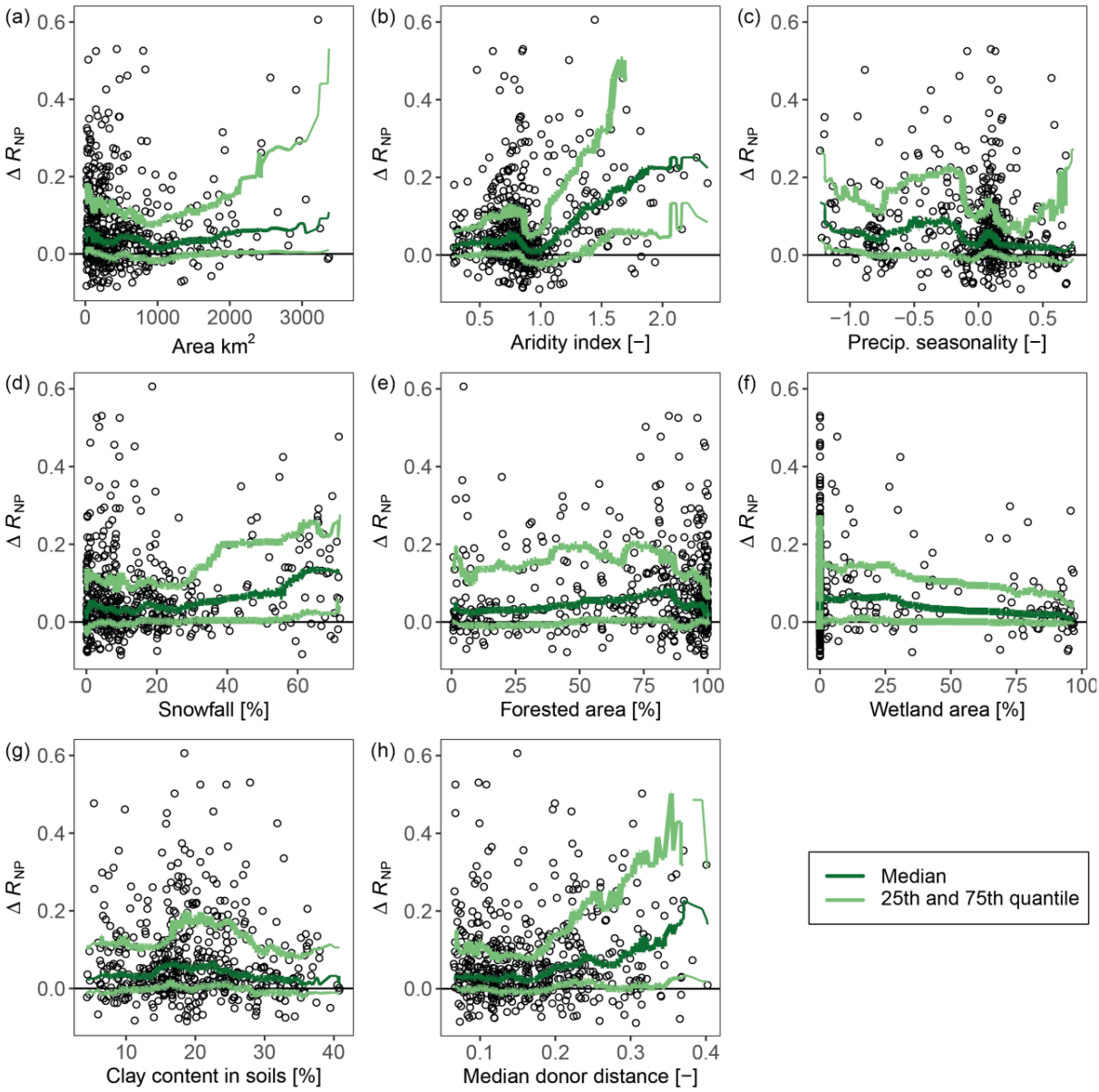


Figure 5. Difference in validation model efficiency (ΔR_{NP}) between a regionalisation with attribute similarity and an informed regionalisation with 24 discharge observations (efficiency of the median sampling year) vs. a-g) catchment attributes and h) the median distance in the attribute space between the 'ungauged' catchment and its donors. Results are presented for all 579 study catchments, whereby positive efficiency difference indicates an increase in prediction efficiency using information of a few discharge observations. The lines represent averaged values of the median, the 25th quantile and the 75th quantile over a moving window of 11 to 101 catchments (smaller moving windows were used at the lower and upper boundaries of the attribute data as indicated by thinner lines towards the boundaries).

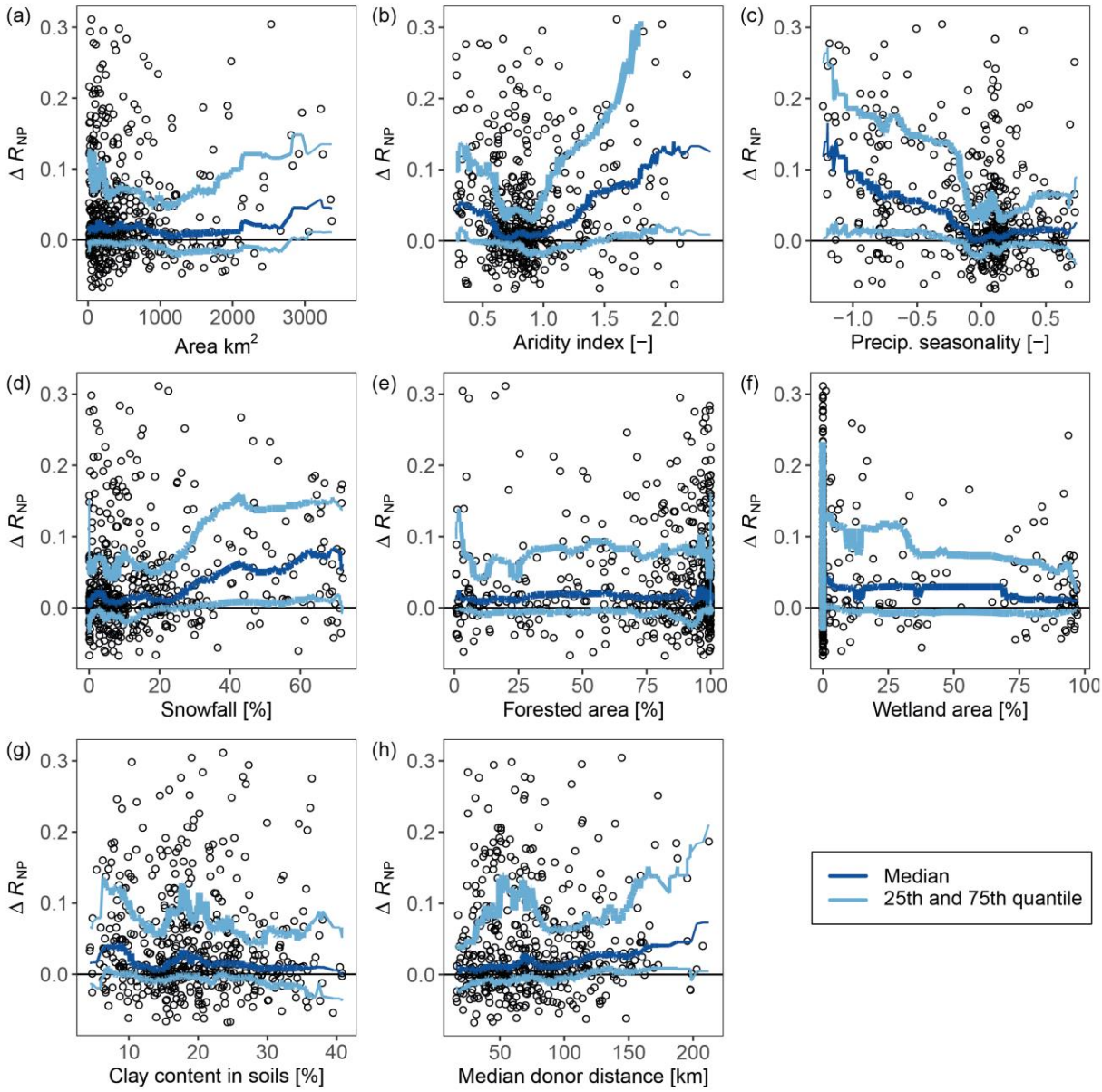


Figure 6. Difference in validation model efficiency (ΔR_{NP}) between a regionalisation with spatial proximity and an informed regionalisation with 24 discharge observations (efficiency of the median sampling year) vs. a-g) catchment attributes and h) the median distance between the 'ungauged' catchment and its donors. Results are presented for all 579 study catchments, whereby positive efficiency difference indicates an increase in prediction efficiency using information of a few discharge observations. The lines represent averaged values of the median, the 25th quantile and the 75th quantile over a moving window of 11 to 101 catchments (smaller moving windows were used at the lower and upper boundaries of the attribute data as indicated by thinner lines towards the boundaries).

5.3 About the information content of different discharge sampling years

We used discharge observations from 10 different years to analyse the effect of a sampling year on the regionalisation. For the case of 24 discharge observations (Fig. 7), the most informative year improved the classical regionalisation with attribute similarity and spatial proximity in 94 % and 92 % of the catchments, respectively. The positive effect of discharge measurements on regionalisation was observed for most sampling years, although the number of catchments experiencing the positive effect steadily decreased with increasing rank number. Only in one (attribute similarity) or two (spatial proximity) out of 10 sampling years, the selected discharge observations were disinformative for the regionalisation of discharge in a majority of catchments.

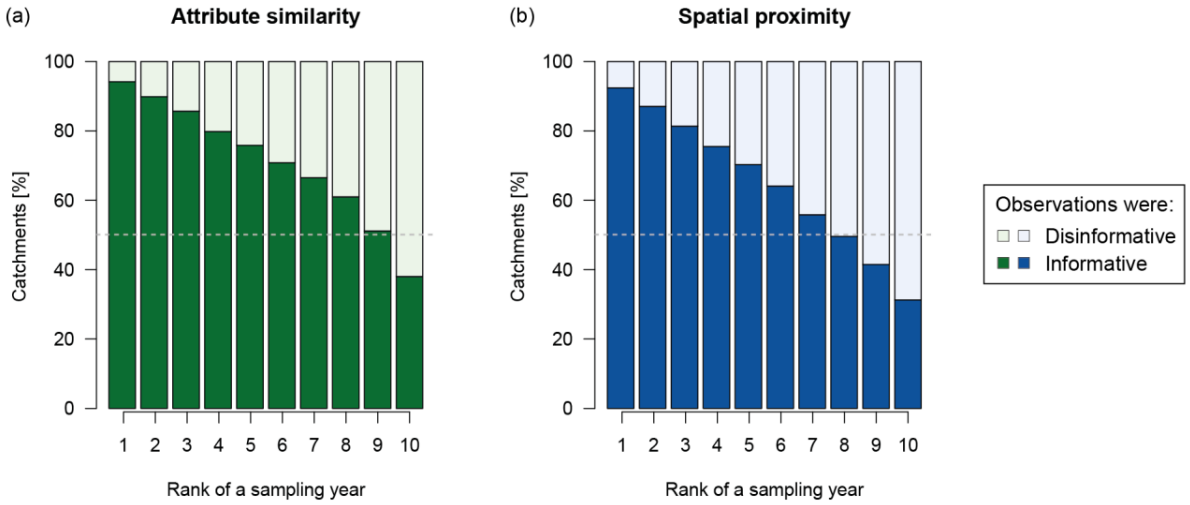


Figure 7. Effect of a sampling year on the value of discharge observations for regionalisation in the 579 study catchments. The 10 sampling years are ranked by the validation model efficiency (R_{NP}) of the informed regionalisation with 24 discharge observations. Results are presented for the informed regionalisation based on a) attribute similarity and b) spatial proximity.

Given that the value of discharge observations varies across years, the question arises "what is a good sampling year?". Table 2 presents the correlation coefficients between the hydrometerological conditions of a sampling year and the corresponding model efficiency in that particular year. Results are only presented for catchment types with significant correlation coefficients although most correlations were still rather weak. For the presented catchment types, regionalisation with attribute similarity was more sensitive to yearly hydroclimatic aspects than regionalisation with spatial proximity. Overall, the magnitude of the highest discharge observation taken in a sampling year had the strongest (negative) effect on model

efficiency among the tested variables, followed by the sum of annual precipitation or winter precipitation. This means that sampling years characterized by exceptionally high peak discharge, or high annual or winter precipitation were the least informative for regionalisation of discharge in arid catchments, snow dominated catchments, and winter-precipitation dominated catchments.

Table 2. Correlation between Model Efficiency and Hydrometerological Conditions in each of the 10 Sampling Years.

		Precipitation sum				Discharge magnitude		
	Catchments	Annual	Winter	Summer	Snowfall	Max.	Median	Range
Attribute similarity	All	**-0.04**	**-0.04**	-0.02	-0.02	**-0.12**	-0.02	-0.02
	Arid	**-0.08**	**-0.13**	0.03	**-0.13**	**-0.20**	**-0.10**	-0.04
	Snowy	**-0.23**	**-0.20**	**-0.11**	-0.05	**-0.24**	**-0.09**	0.03
	Winter precip.	**-0.14**	**-0.15**	-0.05	-0.05	**-0.16**	-0.04	-0.02
Spatial proximity	All	0.00	0.02	**-0.04**	0.01	**-0.09**	0.01	-0.02
	Arid	-0.04	-0.05	-0.02	-0.06	**-0.12**	-0.03	**-0.08**
	Snowy	**-0.12**	-0.06	**-0.14**	0.00	**-0.16**	-0.05	**-0.09**
	Winter precip.	-0.01	-0.01	**-0.05**	0.00	**-0.07**	0.00	**-0.05**

Note: Spearman rank correlation coefficients were calculated for normalized variables and normalized model efficiencies for all catchments (n = 579), arid catchments (aridity index >= 1.2, n = 110), snow dominated catchments (percentage of annual precipitation falling as snow >= 50 %, n = 71), and catchments with predominantly winter precipitation (seasonality index <= -0.2, n = 104). Significant correlations (p-value < 0.05) are marked in bold letters. Model efficiency used for calculating correlations is the difference in validation model efficiency (ΔR_{NP}) between a classical regionalisation with attribute similarity or spatial proximity and an informed regionalisation with 24 discharge observations.

5.4 How many discharge observations are needed?

Figure 8 presents the effect of the number of discharge observations on the regionalisation. An increasing number of observations in the least informative year not only improved efficiencies, but also clearly reduced the variability in model performance between catchments. More importantly, with the use of more observations, a sampling year could change from being mostly disinformative to being informative for a considerable number of catchments. In a median year, median model performance only slightly increased with an increasing number of observations. However, informing regionalisation with 24 instead of 3 observations increased the number of catchments that were better predicted by the informed regionalisation by about 10 %. In the most informative year, 3 discharge observations had a comparable effect on model performance as 24 discharge observations.

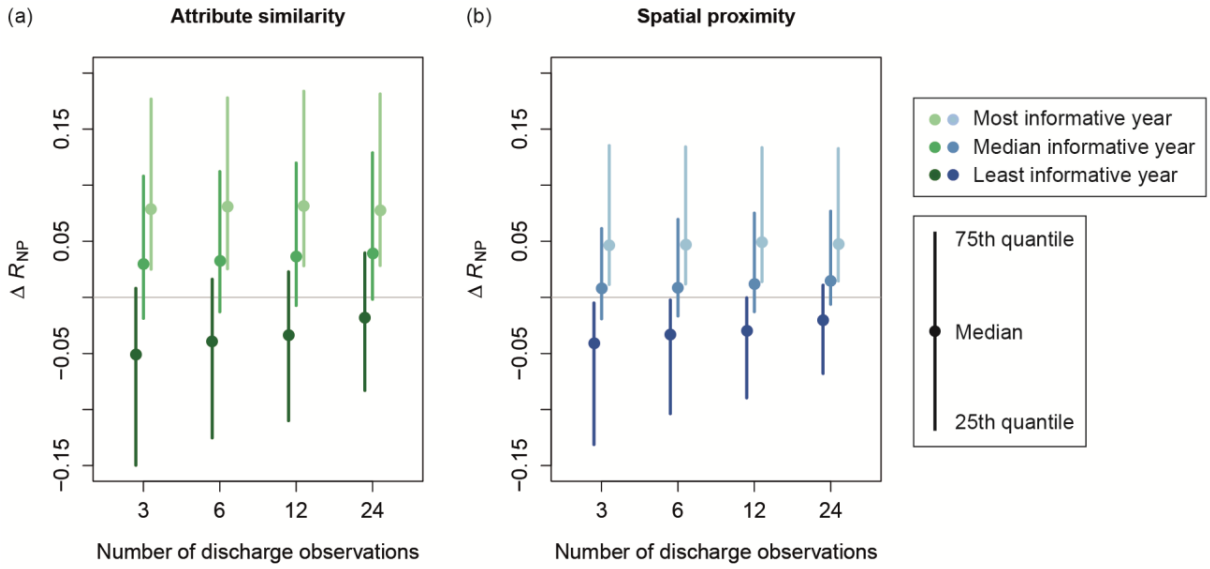


Figure 8. Effect of the number of discharge observation on the difference in validation model efficiency (ΔR_{NP}) between a classical regionalisation and an informed regionalisation in the 579 study catchments. The effect of a variable number of observations on regionalisation with a) attribute similarity and b) spatial proximity is presented for the most informative sampling year, the median sampling year, and the least informative sampling year. Positive efficiency differences (ΔR_{NP}) indicate an increase in prediction efficiency using information of a few discharge observations.

6 Discussion

The result that a limited number of discharge observations can improve predictions in otherwise ungauged catchments is in agreement with Rojas-Serna et al. (2016) and Viviroli and Seibert (2015), who concluded that a few randomly selected discharge observations or a few observations during mean-flow conditions proved to be a valuable source of information beyond classical regionalisation. The value of such observations for regionalisation was generally higher for the attribute-similarity based approach than for the spatial-proximity approach. A possible explanation for this variable value of data is the poorer performance of the attribute-based regionalisation, which leaves more room for improvement, than for regionalisation with spatial proximity. The superior performance of spatial proximity was also observed in the comparative regionalisation studies of Oudin et al. (2008) and Zhang and Chiew (2009). Factors such as a relatively dense streamflow gauging network (Lebecherel et al., 2016; Oudin et al., 2008; Yang et al., 2018) or a suboptimal selection of key catchment attributes (Arsenault & Brissette, 2014; Oudin et al., 2008) could have favoured the spatial proximity approach over regionalisation with attribute similarity. In fact, results of this study showed that donor catchments selected by attribute similarity were up to several hundreds of kilometres away from the ungauged catchment in many semi-arid catchments in the Southwest.

This probably impaired the representativeness of the selected donors for the runoff response in the ungauged catchment. Under such circumstances and given the relatively dense streamflow gauging network in the data set, spatial proximity certainly has the advantage that it implicitly considers relevant attributes influencing major hydrograph aspects. In addition to model efficiency, criteria such as objectivity and reproducibility could also be seen as a benefit of a spatial-proximity based regionalisation approach. While in this study, donor catchments were selected by either spatial proximity or attribute similarity, attempts of combining both approaches have been shown to be a promising approach for the selection of potential donor catchments (Buytaert & Beven, 2009; Oudin et al., 2008; Samuel et al. 2011; Yang et al., 2018; Zhang and Chiew, 2009).

Independent of the classical regionalisation approach, discharge observations were most informative in arid catchments, snow dominated catchments, and winter-precipitation dominated catchments. These catchments generally have a distinct runoff regime with a pronounced high-flow period. The discharge observations selected to inform the regionalisation in this study were sampled during these period of high flow. They therefore provided information for the regionalisation at a time when dominant runoff processes were active. These results are comparable to those of Viviroli and Seibert (2015), who reported that discharge observations during the snowmelt and icemelt season were most valuable for informing regionalisation in snow dominated or glaciated catchments. They furthermore showed that more observations were needed to effectively inform regionalisation for catchments with a predominantly pluvial regime because of the randomness of rain events between and within years as opposed to the reoccurring process of snowmelt and icemelt. The variability in precipitation and the related timing of the dominant runoff processes could also provide an explanation for the limited value of discharge observations found in large parts of the central region of the eastern United States.

The value of discharge observations for the regionalisation varied across sampling years, whereby years with high sums of annual or winter precipitation and therefore relatively high discharge events were the least informative ones. This was unexpected at first, because it has been shown that runoff models could be calibrated with a limited number of discharge observations, especially if they were sampled during wet periods (Melsen et al., 2014; Vrugt et al., 2006; Yapo et al., 1996) or peak events (Pool et al., 2017; Seibert & McDonnell, 2015) when dominant runoff processes were active. However, under unusually wet conditions special runoff processes might govern catchment runoff responses. Using discharge observations taken during such unusual conditions can inform the regionalisation with data of limited representativeness, which ultimately favours parameter sets that reproduce rather exceptional runoff responses. However, it is important to note that the correlations between the value of observations and the hydrometeorological conditions in a sampling year were rather weak and only significant for arid catchments, snow dominated

catchments and winter-precipitation dominated catchments. These results therefore have to be interpreted with some caution. More detailed insights into the value of individual sampling years could be gained by an inductive approach that addresses the influence of large-scale climate phenomena such as El-Niño Southern Oscillation, rating curve uncertainties affecting the value of discharge observations at peak flows, or disinformation at the event level introduced by a mismatch between precipitation input and runoff response (Beven & Westerberg, 2011). However, such an in-depth analysis on the value of sampling years was not conducted within this study.

The sampling year not only influenced model performance, but also affected the number of observations needed to inform regionalisation. Increasing the number of discharge observations strongly improved regionalisation for observations collected in the least informative year, probably because the effect of an individual unusual event could be balanced by additional and more representative observations. In contrast, the characteristic runoff response could be captured by as few as three observations if these observations were collected in the most informative sampling year.

The results of this study are based on the strategic extraction of a few discharge observations from the observed time series of catchments and therefore provide insights in what could be achieved at best. In practice, decisions on the number of observations, the dates of observations, or the sampling year may be restricted by economical or organizational factors. Cost-benefit analysis for real case studies could be a way to bridge the gap between the theoretical and practical value of a limited number of discharge observations for the prediction in ungauged catchments.

An additional practical limitation of this study is the fact that discharge observations used to inform regionalisation correspond to mean daily values, whereas observations collected during field campaigns are almost instantaneous. The differences between instantaneous discharge values (reported discharge data at 15 minutes interval) and mean daily discharge were looked at for eight catchments representing the typical range of catchment areas (10 km^2, 100 km^2, 1000 km^2, and 10 000 km^2) encountered in the CAMELS data set. Thereby it could be observed that mean daily discharge can deviate considerably from instantaneous discharge observations at days with peak flows. However, instantaneous measurements can be regarded as representative for the mean daily values during most other periods including event recessions and low flow periods when within-day flow variations are relatively small. Similarly, a more detailed analysis of the value of sub-daily discharge observations by Viviroli and Seibert (2015) indicated a limited added value of several subsequent instantaneous discharge observations within a few hours.

7 Conclusions

Many catchments lack continuous discharge time series and the prediction of discharge relies on regionalisation. However, it might still be possible to collect a limited number of discharge observations during short field campaigns. In this study, we evaluated the value of such a limited number of discharge observations for informing parameter regionalisation using a large-sample data set of the United States. Results demonstrated that a few discharge observations improved regionalisation in the majority of catchments and that observations were especially effective in arid catchments, snow dominated catchments, and winter-precipitation dominated catchments. Discharge observations from years with moderate to low peak-flow magnitudes were the most informative ones, whereby 3 observations could be of comparable value as 24 observations if collected in these most informative years. The results demonstrate the value of a small number of streamflow observations and indicate that short field campaigns can improve the basis for decision making in ungauged basins.

Acknowledgments and Data

This work was supported by the University of Zurich. Hydrometeorological data, catchment attributes and catchment boundaries were made available by Addor et al. (2017) and Newman et al. (2015). SRTM elevation data was used from Jarvis et al. (2008) and data on the distribution of wetlands was extracted from the data set of Lehner and Döll (2004).

References

Addor, N., Newman, A. J., Mizukami, N., & Clark, M. P. (2017). The CAMELS data set: Catchment attributes and meteorology for large-sample studies. *Hydrology and Earth System Sciences*, *21*(10), 5293–5313. https://doi.org/10.5194/hess-21-5293-2017

Arsenault, R., & Brissette, F. (2014). Continuous streamflow prediction in ungauged basins: The effects of equifinality and parameter set selection on uncertainty in regionalization approaches. *Water Resources Research*, (50), 6135–6153. https://doi.org/10.1002/2013WR014898

Bárdossy, A. (2007). Calibration of hydrological model parameters for ungauged catchments. *Hydrology and Earth System Sciences*, *11*(2), 703–710. https://doi.org/10.5194/hess-11-703-2007

Bergström, S. (1976). *Development and application of a conceptual runoff model for Scandinavian catchments*. Norrköping, Sweden: SMHI.

Beven, K., & Freer, J. (2001). Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology*,

249(1–4), 11–29. https://doi.org/10.1016/S0022-1694(01)00421-8

Beven, K., & Westerberg, I. (2011). On red herrings and real herrings: Disinformation and information in hydrological inference. *Hydrological Processes*, *25*(10), 1676–1680. https://doi.org/10.1002/hyp.7963

Blöschl, G., & Sivapalan, M. (1995). Scale issues in hydrological modelling: A review. *Hydrological Processes*, *9*, 251–290. https://doi.org/10.5194/hess-19-4559-2015

Burn, D. H. (1990). Evaluation of regional flood frequency analysis with a region of influence approach. *Water Resources Research*, *26*(10), 2257–2266. https://doi.org/10.1029/90WR01192

Burn, D. H., & Boorman, D. B. (1993). Estimation of hydrological parameters at ungauged catchments. *Journal of Hydrology*, *143*(3–4), 429–454. https://doi.org/10.1016/0022-1694(93)90203-L

Buytaert, W., & Beven, K. (2009). Regionalization as a learning process. *Water Resources Research*, *45*(11), W11419. https://doi.org/10.1029/2008WR007359

Gupta, H. V., Kling, H., Yilmaz, K. K., & Martinez, G. F. (2009). Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *Journal of Hydrology*, *377*(1–2), 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Harlin, J. (1991). Development of a process oriented calibration scheme for the HBV hydrological model. *Nordic Hydrology*, *22*, 15–36. https://doi.org/10.2166/nh.1991.002

He, Y., Bárdossy, A., & Zehe, E. (2011). A review of regionalisation for continuous streamflow simulation. *Hydrology and Earth System Sciences*, *15*(11), 3539–3553. https://doi.org/10.5194/hess-15-3539-2011

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., et al. (2013). A decade of Predictions in Ungauged Basins (PUB) - A review. *Hydrological Sciences Journal*, *58*(6), 1198–1255. https://doi.org/10.1080/02626667.2013.803183

Jarvis, A., Reuter, H., Nelson, A., & Guevara, E. (2008). Hole-filled SRTM for the globe Version 4, available from the CGIAR-CSI SRTM 90m. Retrieved from http://srtm.csi.cgiar.org

Kokkonen, T. S., Jakeman, A. J., Young, P. C., & Koivusalo, H. J. (2003). Predicting daily flows in ungauged catchments: Model regionalization from catchment descriptors at the Coweeta Hydrologic Laboratory, North Carolina. *Hydrological Processes*, *17*(11), 2219–2238. https://doi.org/10.1002/hyp.1329

Lebecherel, L., Andréassian, V., & Perrin, C. (2016). On evaluating the robustness of spatial-proximity-based regionalization methods. *Journal of Hydrology*, *539*, 196–203. https://doi.org/10.1016/j.jhydrol.2016.05.031

Lehner, B., & Döll, P. (2004). Development and validation of a global database of lakes, reservoirs and wetlands. *Journal of Hydrology*, *296*(1–4), 1–22. https://doi.org/10.1016/j.jhydrol.2004.03.028

Lindström, G., Johansson, B., Persson, M., Gardelin, M., & Bergström, S. (1997). Development and test of the distributed HBV-96 hydrological model. *Journal of Hydrology*, *201*(1–4), 272–288. https://doi.org/10.1016/S0022-1694(97)00041-3

McIntyre, N., Lee, H., Wheater, H., Young, A., & Wagener, T. (2005). Ensemble predictions of runoff in ungauged catchments. *Water Resources Research*, *41*(12), 1–14. https://doi.org/10.1029/2005WR004289

Melsen, L. A. LA, Teuling, A. J., van Berkum, S. W. S., Torfs, P. J. J. F., & Uijlenhoet, R. (2014). Catchments as simple dynamical systems: A case study on methods and data requirements for parameter identification. *Water Resources Research*, *50*, 5577–5596. https://doi.org/10.1002/2013WR014720

Merz, R., & Blöschl, G. (2004). Regionalisation of catchment model parameters. *Journal of Hydrology*, *287*(1–4), 95–123. https://doi.org/10.1016/j.jhydrol.2003.09.028

Milligan, G. W., & Cooper, M. C. (1988). A study of standardizaction of variables in cluster analysis. *Journal of Classification*, *5*, 181–204. https://doi.org/10.1007/BF01897163

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Oudin, L., Andréassian, V., Perrin, C., Michel, C., & Le Moine, N. (2008). Spatial proximity, physical similarity, regression and ungaged catchments: A comparison of regionalization approaches based on 913 French catchments. *Water Resources Research*, *44*(3), 1–15. https://doi.org/10.1029/2007WR006240

Parajka, J., Viglione, A., Rogger, M., Salinas, J. L., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins – Part 1: Runoff-hydrograph studies. *Hydrology and Earth System Sciences*, *17*(5), 1783–1795. https://doi.org/10.5194/hess-17-1783-2013

Pool, S., Vis, M., & Seibert, J. (n.d.). Evaluating model performance: a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*.

Pool, S., Viviroli, D., & Seibert, J. (2017). Prediction of hydrographs and flow-duration curves in almost ungauged catchments: Which runoff measurements are most informative for model calibration? *Journal of Hydrology*, *554*, 613–622. https://doi.org/10.1016/j.jhydrol.2017.09.037

Priestley, C. H. B., & Taylor, R. J. (1972). On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters. *Monthly Weather Review*, *100*(2), 81–92. https://doi.org/10.1175/1520-0493(1972)100<0081:OTAOSH>2.3.CO;2

Razavi, T., Coulibaly, P., & Asce, M. (2013). Streamflow prediction in ungauged basins: Review of regionalization methods. *Journal of Hydrologic Engineering*, *18*(8), 958–975. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000690

Rojas-Serna, C., Lebecherel, L., Perrin, C., Andréassian, V., & Oudin, L. (2016). How should a rainfall-runoff model be parameterized in an almost ungauged catchment? A methodology tested on 609 catchments. *Water Resources Research*, *52*, 4765–4784. https://doi.org/10.1002/ 2015WR018549

Samuel, J., Coulibaly, P., & Metcalfe, R. A. (2011). Estimation of continuous streamflow in Ontario ungauged basins: Comparison of regionalization methods. *Journal of Hydrologic Engineering*, *16*(5), 447–459. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000338

Seibert, J. (2000). Multi-criteria calibration of a conceptual runoff model using a genetic algorithm. *Hydrology and Earth System Sciences*, *4*(2), 215–224. https://doi.org/10.5194/hess-4-215-2000

Seibert, J., & Beven, K. J. (2009). Gauging the ungauged basin: How many discharge measurements are needed? *Hydrology and Earth System Sciences*, *13*(6), 883–892. https://doi.org/10.5194/hess-13-883-2009

Seibert, J., & McDonnell, J. J. (2015). Gauging the ungauged basin: Relative value of soft and hard data. *Journal of Hydrologic Engineering*, *20*(1), A4014004. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000861.

Seibert, J., & Vis, M. J. P. (2012). Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrology and Earth System Sciences*, *16*(9), 3315–3325. https://doi.org/10.5194/hess-16-3315-2012

Seibert, J., Vis, M. J. P., Lewis, E., & van Meerveld, H. J. (2018). Upper and lower benchmarks in hydrological modelling. *Hydrological Processes*, *32*(8), 1120–1125.

https://doi.org/10.1002/hyp.11476

Singh, S. K., & Bárdossy, A. (2012). Calibration of hydrological models on hydrologically unusual events. *Advances in Water Resources*, *38*, 81–91. https://doi.org/10.1016/j.advwatres.2011.12.006

Sivapalan, M., Takeuchi, K., Franks, S. W., Gupta, V. K., Karambiri, H., Lakshmi, V., et al. (2003). IAHS Decade on Predictions in Ungauged Basins (PUB), 2003–2012: Shaping an exciting future for the hydrological sciences. *Hydrological Sciences Journal*, *48*(6), 857–880. https://doi.org/10.1623/hysj.48.6.857.51421

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit Region. *Economic Geography*, *46*, 234–240. https://doi.org/10.1126/science.11.277.620

Viviroli, D., & Seibert, J. (2015). Can a regionalized model parameterisation be improved with a limited number of runoff measurements? *Journal of Hydrology*, *529*, 49–61. https://doi.org/10.1016/j.jhydrol.2015.07.009

Viviroli, D., Mittelbach, H., Gurtz, J., & Weingartner, R. (2009). Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland–Part II: Parameter regionalisation and flood estimation results. *Journal of Hydrology*, *377*(1), 208–225. https://doi.org/10.1016/j.jhydrol.2009.08.022

Vrugt, J. A., Gupta, H. V., Dekker, S. C., Sorooshian, S., Wagener, T., & Bouten, W. (2006). Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting model. *Journal of Hydrology*, *325*(1–4), 288–307. https://doi.org/10.1016/j.jhydrol.2005.10.041

Yang, X., Magnusson, J., Rizzi, J., & Xu, C.-Y. (2018). Runoff prediction in ungauged catchments in Norway: Comparison of regionalization approaches. *Hydrology Research*, *49*(2), 487–505. https://doi.org/10.2166/nh.2017.071

Yapo, P. O., Gupta, H. V., & Sorooshian, S. (1996). Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data. *Journal of Hydrology*, *181*(1–4), 23–48. https://doi.org/10.1016/0022-1694(95)02918-4

Zhang, Y., & Chiew, F. H. S. (2009). Relative merits of different methods for runoff predictions in ungauged catchments. *Water Resources Research*, *45*, W07412. https://doi.org/10.1029/2008WR007504